BIONTECH | InstaDeep

# AI Day

October 1, 2025

# This slide presentation includes forward-looking statements

This presentation contains forward-looking statements within the meaning of the Private Securities Litigation Reform Act of 1995, as amended. In some cases, forward-looking statements can be identified by terminology such as “will,” “may,” “should,” “expects,” “intends,” “plans,” “aims,” “anticipates,” “believes,” “estimates,” “predicts,” “potential,” “continue,” or the negative of these terms or other comparable terminology, although not all forward-looking statements contain these words. The forward-looking statements in this presentation are neither promises nor guarantees, and you should not place undue reliance on these forward-looking statements because they involve known and unknown risks, uncertainties, and other factors, many of which are beyond BioNTech’s control; and which could cause actual results to differ materially from those expressed or implied by these forward-looking statements. You should review the risks and uncertainties described under the heading “Risk Factors” in BioNTech's Quarterly Report on Form 6-K for the period ended June 30, 2025; and in subsequent filings made by BioNTech with the SEC, which are available on the SEC’s website at https://www.sec.gov/. Except as required by law, BioNTech disclaims any intention or responsibility for updating or revising any forward-looking statements contained in this presentation in the event of new information, future developments or otherwise. These forward-looking statements are based on BioNTech’s current expectations and speak only as of the date hereof.

Furthermore, certain statements contained in this presentation relate to or are based on studies, publications, surveys and other data obtained from third-party sources and BioNTech’s own internal estimates and research. While BioNTech believes these third-party sources to be reliable as of the date of this presentation, it has not independently verified, and makes no representation as to the adequacy, fairness, accuracy or completeness of, any information obtained from third-party sources. In addition, any market data included in this presentation involves assumptions and limitations, and there can be no guarantee as to the accuracy or reliability of such assumptions. While BioNTech believes its own internal research is reliable, such research has not been verified by any independent source. This presentation contains references to our trademarks and to trademarks belong to other entities. Solely for convenience, trademarks and trade names referred to, including logos, artwork and other visual displays, may appear without the ® or TM symbols, but such references are not intended to indicate, in any way, that their respective owners will not assert, to the fullest extent under applicable law, their rights thereto. We do not intend our use or display of other companies’ trade names or trademarks to imply a relationship with, or endorsement or sponsorship of us by, any other companies.

# Agenda

## BioNTech – Building a global immunotherapy powerhouse translating science into survival

| | | |
|---|---|---|
| 14:00 | Advancing a disruptive tech-bio company | Prof. U. Sahin, M.D. |
| 14:15 | Developing the future of AI at BioNTech | K. Beguir |

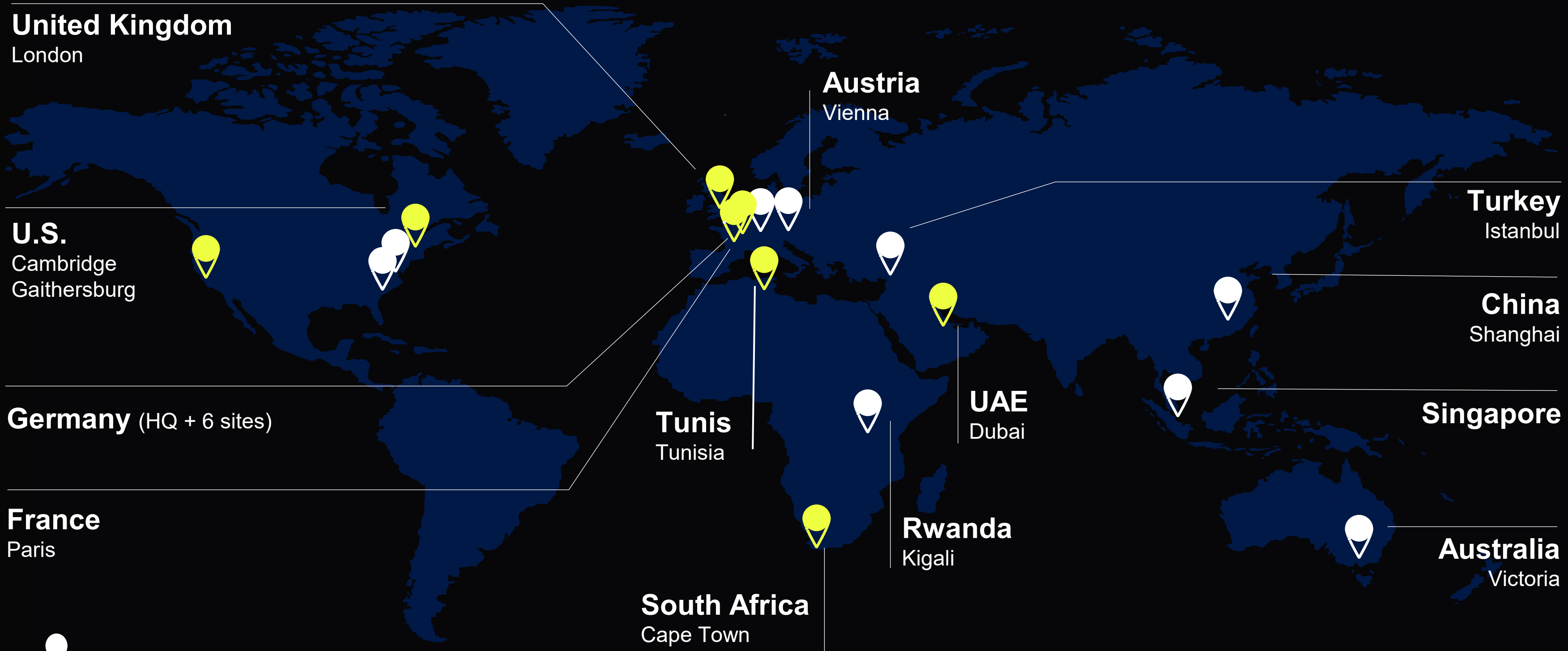
## InstaDeep – Delivering across the full AI stack

| | | |
|---|---|---|
| 14:25 | Compute & model scaling | A. Laterre |
| 14:35 | AI innovation | B. Almeida, B. Guloglu |
| 15:00 | Data acquisition & refinement | N. Lopez Carranza, Y. Ben Dhieb |
| 15:20 | Applications | C. Zhang, L. Walls, A. Delaunay, M. Rooney |
| 15:40 | Audience Q&A | Prof. U. Sahin, M.D., K. Beguir |

# Advancing a disruptive tech-bio company

Ugur Sahin
Founder & CEO
**BioNTech**

# BioNTech's AI capabilities with worldwide reach



**United Kingdom**
London

**U.S.**
Cambridge
Gaithersburg

**Germany** (HQ + 6 sites)

**France**
Paris

**Austria**
Vienna

**Tunis**
Tunisia

**South Africa**
Cape Town

**Rwanda**
Kigali

**UAE**
Dubai

**Turkey**
Istanbul

**China**
Shanghai

**Singapore**

**Australia**
Victoria

○ BioNTech locations

● BioNTech AI centers

# BioNTech – disruptive tech-bio company with pioneering technologies developed through full AI integration

## Multiplatform oncology company

**16** Clinical programs

**>20** Ongoing Phase 2 or 3 trials

REGENERON Genmab DualityBio MediLink Therapeutics Bristol Myers Squibb Genentech OncoC4

## Infectious diseases pipeline

**7** **Clinical programs** in high unmet need indications

Pfizer

## COVID-19 vaccine global impact

**5** Billion doses distributed

## Leader in integrated AI capabilities

InstaDeep®

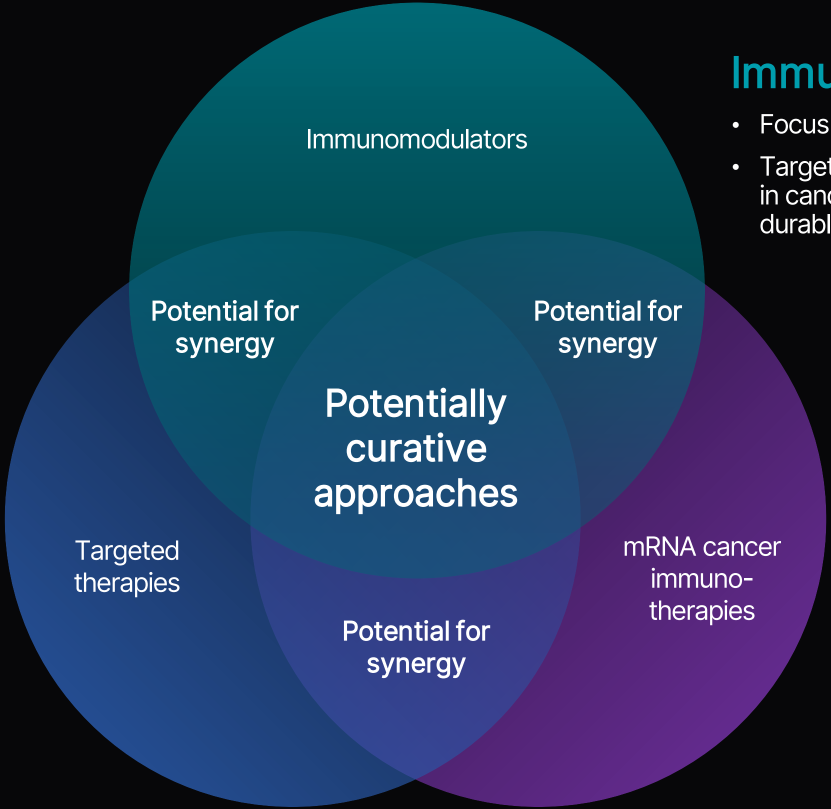
## In-house manufacturing

**4** **Platforms** including individualized mRNA and bispecific antibodies

## Vision

# Building a global immunotherapy powerhouse translating science into survival

# We are uniquely positioned to combine approaches to transform cancer care



## Immunomodulators

- Focus on the critical IO pathways
- Targeting different complementary pathways in cancer immunity cycle may promote a durable anti-tumor effect

## Targeted therapies

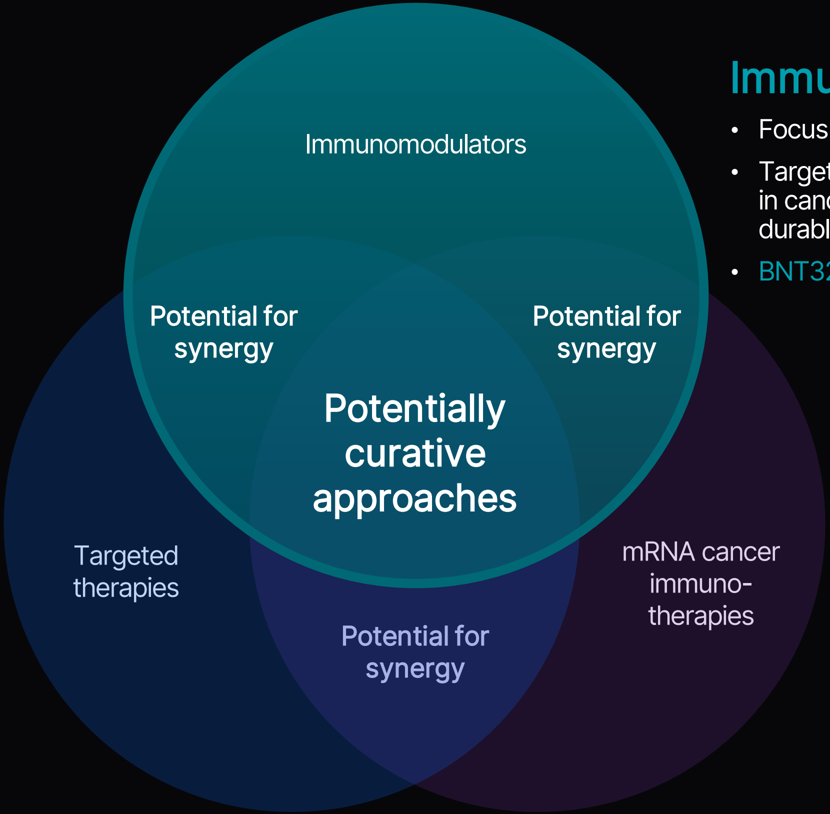
- Precise and potent modalities for fast onset tumor reduction
- ADC as potential "augmenters" of immunomodulators and mRNA cancer immunotherapies
- Focus on HER2, HER3, TROP2, B7H3 ADCs as combination partners

## mRNA cancer immunotherapies

- Eliminate polyclonal residual disease with multi-antigen and individualized approaches
- Polyspecific activity by targeting multiple antigens at once
- Establish long-lasting immunological memory to prevent relapses

ADC = antibody-drug conjugate.

# We are uniquely positioned to combine approaches to transform cancer care



Immunomodulators

Potential for synergy

Potential for synergy

Potentially curative approaches

Targeted therapies

mRNA cancer immuno-therapies

Potential for synergy

## Immunomodulators

- Focus on the critical IO pathways
- Targeting different complementary pathways in cancer immunity cycle may promote a durable anti-tumor effect
- BNT327 pumitamig

## Targeted therapies

- Precise and potent modalities for fast onset tumor reduction
- ADC as potential "augmenters" of immunomodulators and mRNA cancer immunotherapies
- Focus on HER2, HER3, TROP2, B7H3 ADCs as combination partners

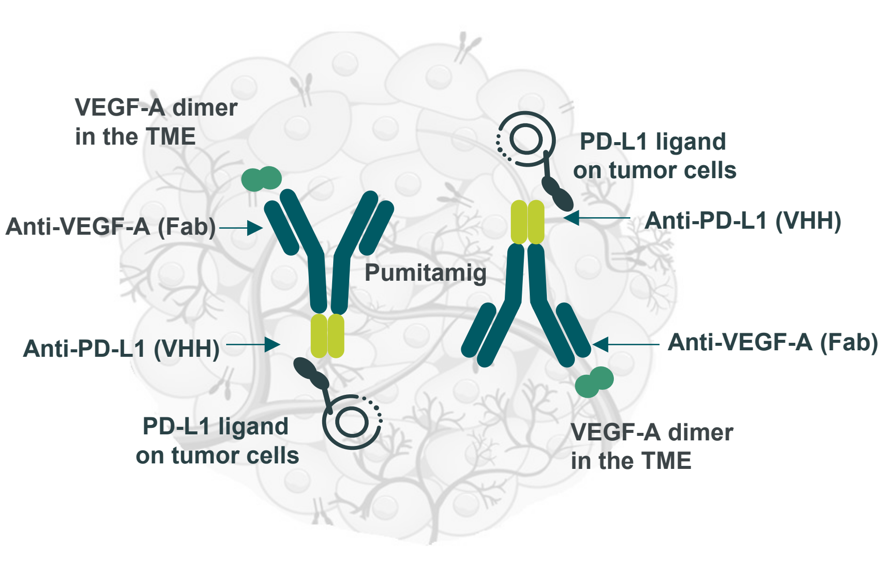
## mRNA cancer immunotherapies

- Eliminate polyclonal residual disease with multi-antigen and individualized approaches
- Polyspecific activity by targeting multiple antigens at once
- Establish long-lasting immunological memory to prevent relapses

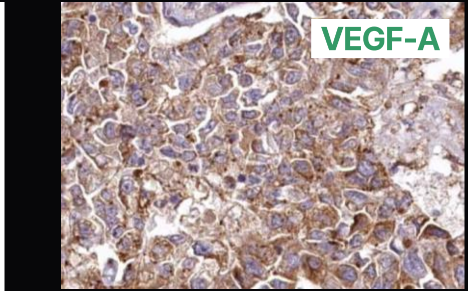
ADC = antibody-drug conjugate.

# Pumitamig's synergistic targeting of PD-L1 and VEGF[1]

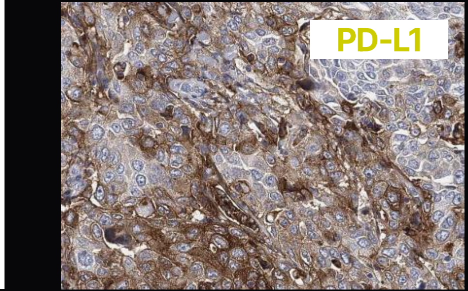
## Tumor microenvironment (TME)



## NSCLC IHC[2]



Local neutralization of angiogenic and immunosuppressive VEGF-A effects



Targeting the TME and blockade of PD-1/PD-L1 signaling

[1] Partnered with Bristol Myers Squibb; [2] IHC data: Human Protein Atlas

# Next-generation bispecific can potentially expand the reach of IO therapy

## Anti-PD-(L)1 approved

Anti-PD-(L)1 therapy addresses

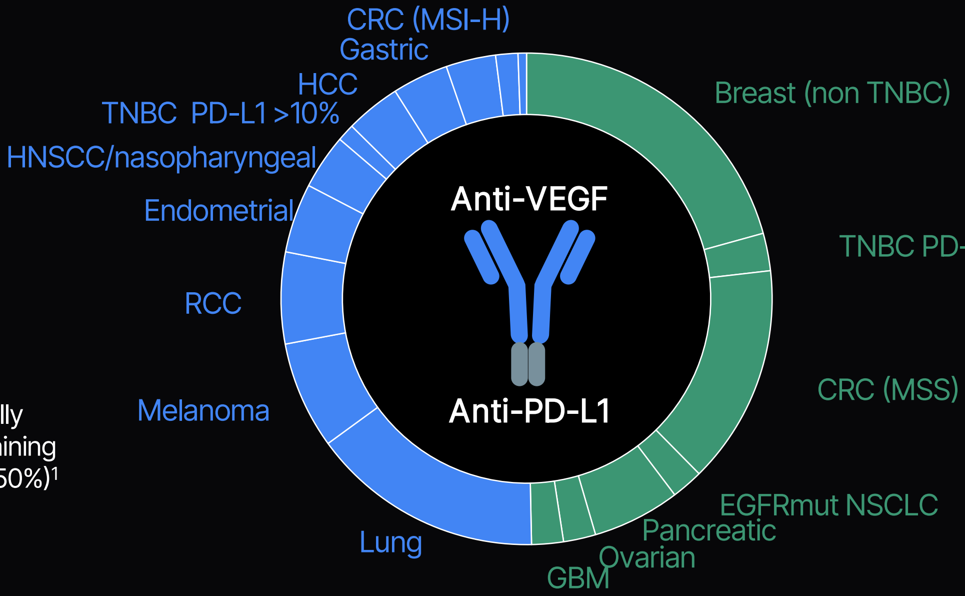
**~1.5 M**

new cancer cases in the US and EU annually with medical need remaining high (5-year survival < 50%)[1]

## Anti-PD-(L)1 not approved

**>1.4 M**

estimated new cancer cases in the US and EU annually that cannot be addressed by current IO therapies

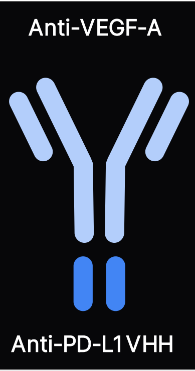


US and EU cancer incidence[2]

[1] NCI SEER https://training.seer.cancer.gov/index.html. 2.US incidence source: NIH and American Cancer Society data EU incidence source: European Cancer Information System

# Landmark strategic collaboration with BMS to advance pumitamig[1]

BIONTECH | Bristol Myers Squibb



## Maximizing potential of next-generation immunomodulator pumitamig[1] with global co-development and co-commercialization partnership

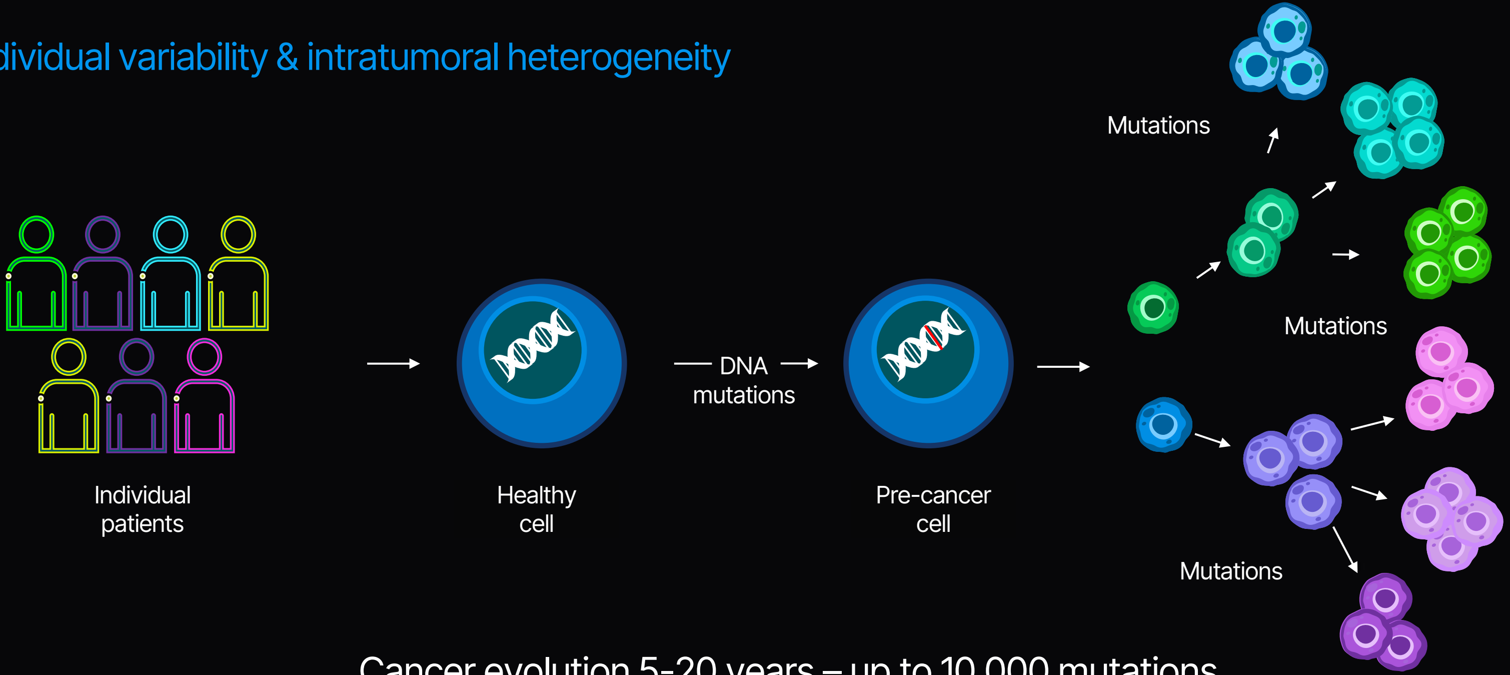
- Bispecific antibody targeting PD-L1 and VEGF-A
- Over 1,200 patients treated in clinical trials across multiple tumor types
- Broad development ongoing in 10+ indications, including initial registrational trials

**Potential to transform standard of care and establish new IO backbone treatment option for patients with high unmet medical needs**

[1] Partnered with Bristol Myers Squibb.

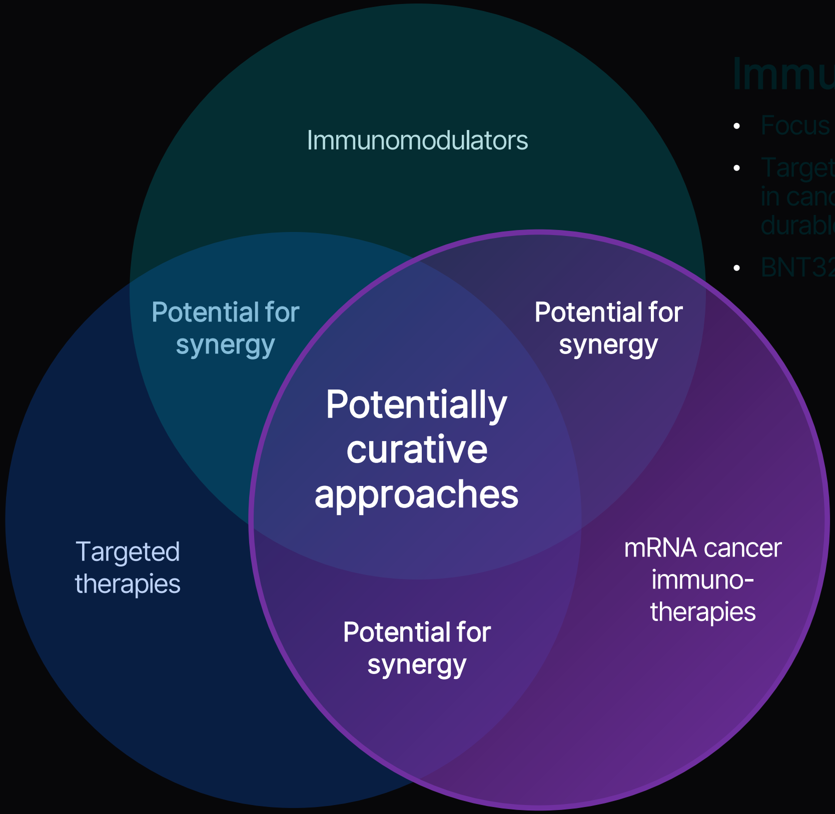
# Root cause of cancer treatment failure

## Interindividual variability & intratumoral heterogeneity



Cancer evolution 5-20 years – up to 10.000 mutations

# We are uniquely positioned to combine approaches to transform cancer care



## Immunomodulators

- Focus on the critical IO pathways
- Targeting different complementary pathways in cancer immunity cycle may promote a durable anti-tumor effect
- BNT327 pumitamig

## Targeted therapies

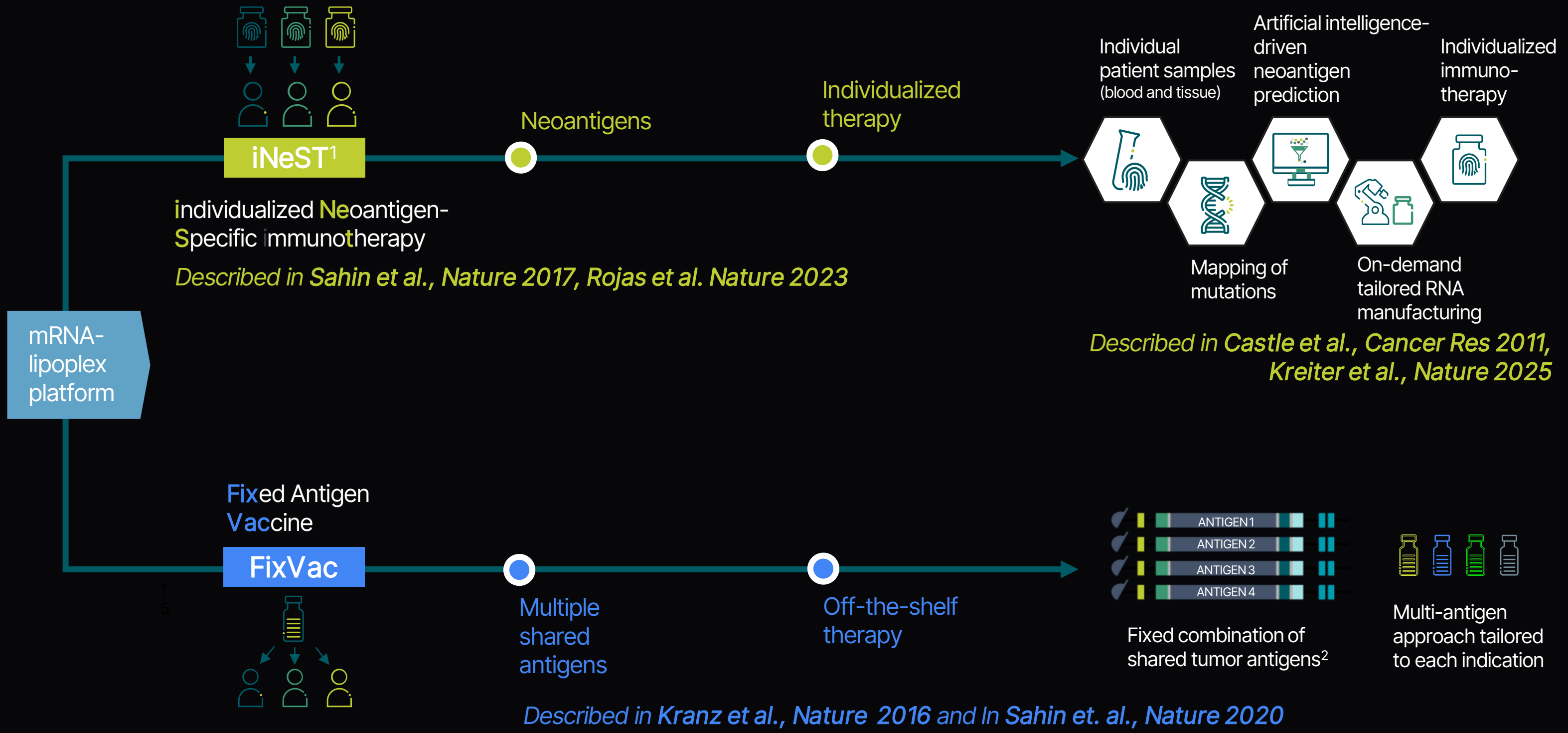
- Precise and potent modalities for fast onset tumor reduction
- ADC as potential “augmenters” of immunomodulators and mRNA cancer immunotherapies
- Focus on HER2, HER3, TROP2, B7H3 ADCs as combination partners

## mRNA cancer immunotherapies

- Eliminate polyclonal residual disease with multi-antigen and individualized approaches
- Polyspecific activity by targeting multiple antigens at once
- Establish long-lasting immunological memory to prevent relapses

ADC = antibody-drug conjugate.

# Leveraging our leadership in mRNA to fully exploit cancer immunotherapy target space with two approaches



mRNA-lipoplex platform

iNeST[1]

individualized Neoantigen-Specific immunotherapy

Neoantigens

Individualized therapy

*Described in Sahin et al., Nature 2017, Rojas et al. Nature 2023*

Individual patient samples (blood and tissue)

Mapping of mutations

Artificial intelligence-driven neoantigen prediction

On-demand tailored RNA manufacturing

Individualized immuno-therapy

*Described in Castle et al., Cancer Res 2011, Kreiter et al., Nature 2025*

Fixed Antigen Vaccine

FixVac

Multiple shared antigens

Off-the-shelf therapy

ANTIGEN 1
ANTIGEN 2
ANTIGEN 3
ANTIGEN 4

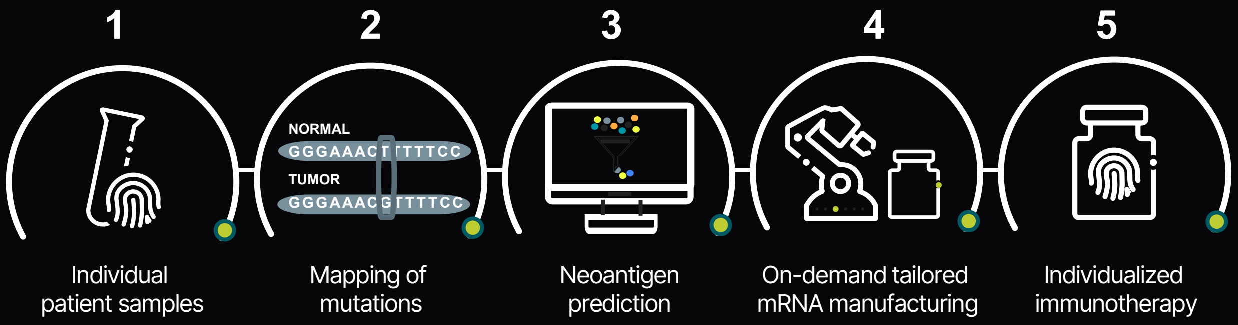
Fixed combination of shared tumor antigens[2]

Multi-antigen approach tailored to each indication

*Described in Kranz et al., Nature 2016 and In Sahin et. al., Nature 2020*

1. Partnered with Genentech, a member of the Roche Group. 2 Antigens vary across programs; 3. T-cell responses analyzed by *ex vivo* multimer staining analysis in blood.

# iNeST: autogene cevumeran driving continuous innovation with data



1. Individual patient samples
2. Mapping of mutations
3. Neoantigen prediction
4. On-demand tailored mRNA manufacturing
5. Individualized immunotherapy

## Driven by data

Potential for continued improvement as more data are generated and analyzed

## Selection algorithms

AI and ML optimization

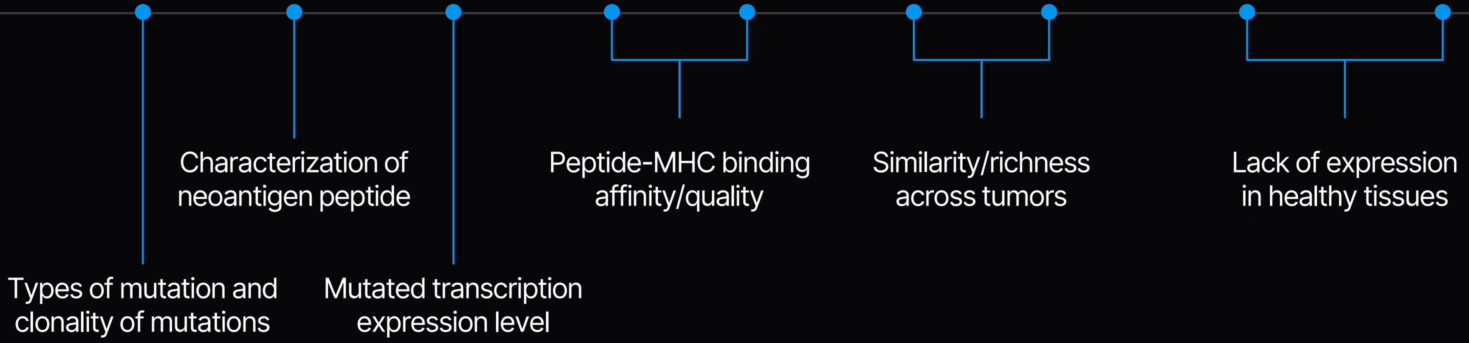
## Just-in-time manufacturing

Dedicated mRNA GMP production facilities

iNeST is being developed in collaboration with Genentech, a member of the Roche Group. Autogene cevumeran is an investigational candidate.

# Neoantigen prediction: how do we identify, predict, and characterize neoantigens?

| Neoantigen rank | Gene | Mutation | Length (aa) | Transcript VAF | MHC I score | MHC II score | Coverage in tumor | VAF in tumor | Coverage in normal tissue | VAF in normal tissue |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | SNF8 | V183M | 27 | 16.05 | 0.1 | 2.16 | 155 | 0.33 | 119 | 0.00 |
| 2 | SEMA7A | G340S | 27 | 1.44 | 0.04 | 8.6 | 113 | 0.44 | 120 | 0.01 |
| 3 | DUS4L | S305P | 26 | 2.07 | 0.28 | 8.54 | 213 | 0.48 | 150 | 0.00 |
| ↓ | | | | | | | | | | |
| 20 | | | | | | | | | | |



Characterization of neoantigen peptide

Peptide-MHC binding affinity/quality

Similarity/richness across tumors

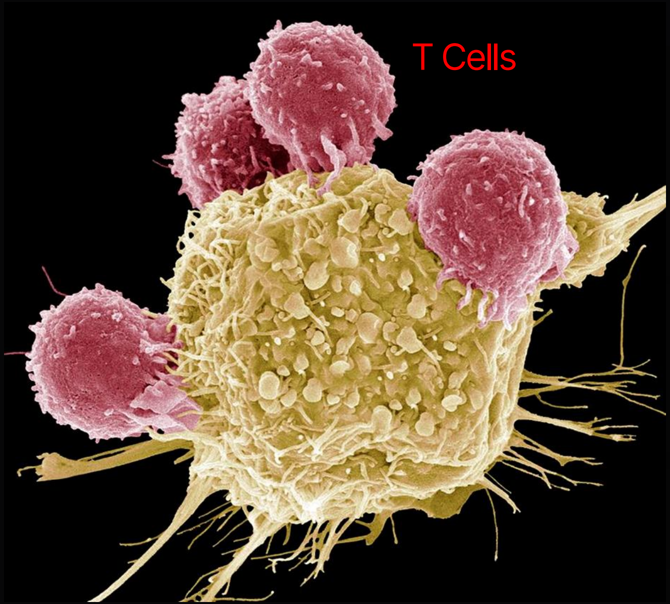
Lack of expression in healthy tissues

Types of mutation and clonality of mutations

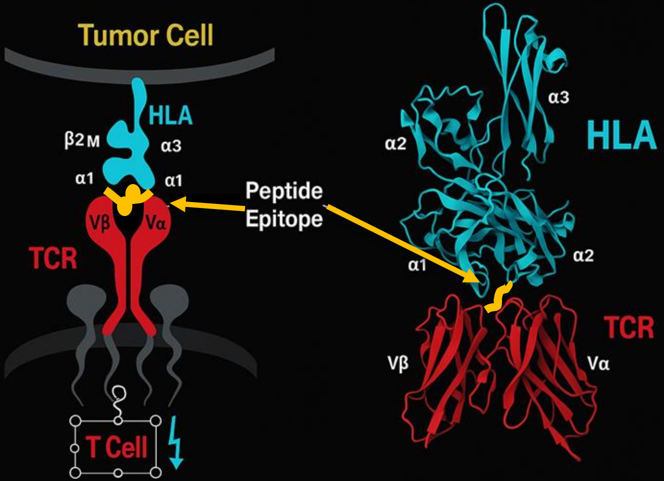
Mutated transcription expression level

Representative data.

# Defining the complex TCR-tumor antigen interaction is an unsolved computational problem



Tumor Cell

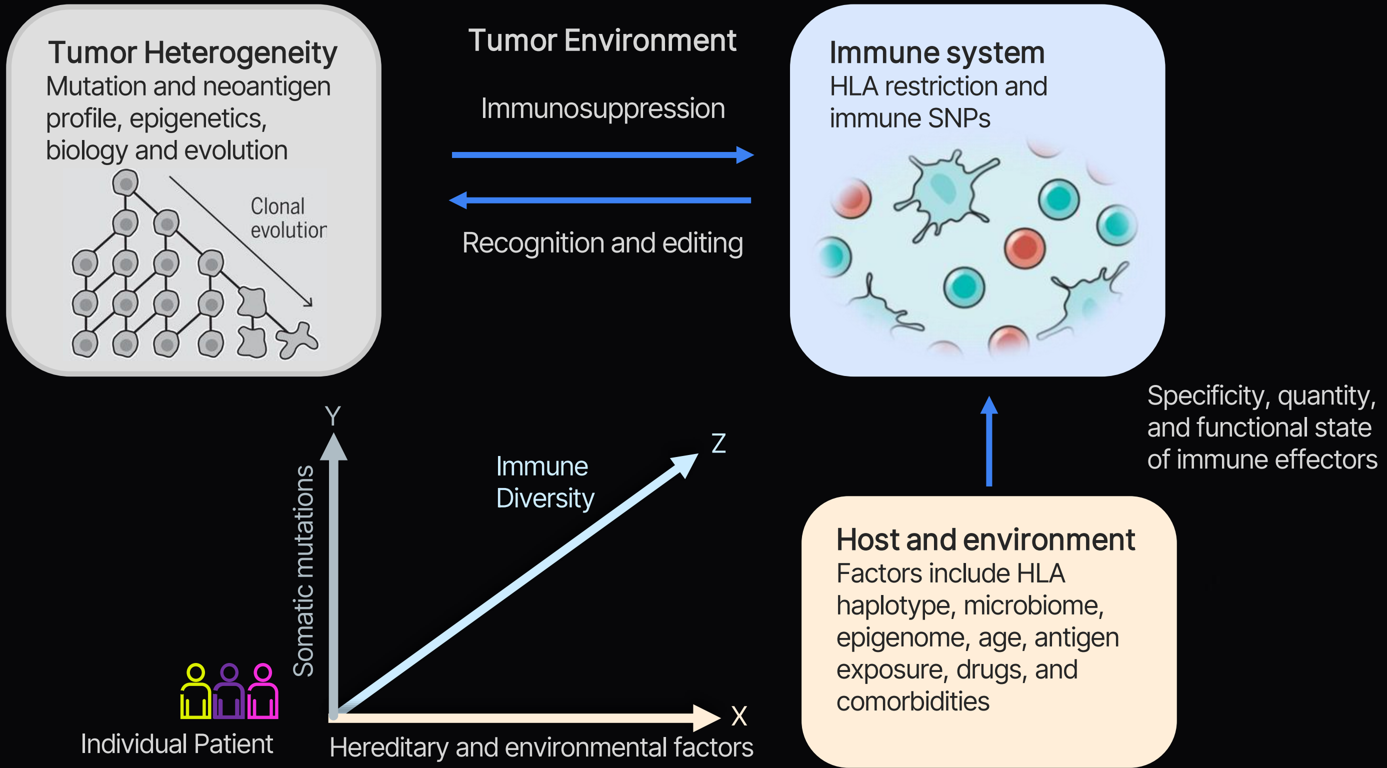


HLA Allele Diversity > 30,000

Peptide Diversity > 100,000

TCR Diversity > $5\times10^8$

# Dimensions of cancer heterogeneity



**Tumor Heterogeneity**
Mutation and neoantigen profile, epigenetics, biology and evolution

**Tumor Environment**

Immunosuppression

Recognition and editing

**Immune system**
HLA restriction and immune SNPs

Specificity, quantity, and functional state of immune effectors

Somatic mutations

Immune Diversity

Hereditary and environmental factors

Individual Patient

**Host and environment**
Factors include HLA haplotype, microbiome, epigenome, age, antigen exposure, drugs, and comorbidities

*Sahin & Türeci, Science 2018*

# Our leading scientific capabilities are fueled by AI to pioneer personalized immunotherapies

## Personalized immunotherapy

- iNeST[1]: Personalized immunotherapy platform levering AI to create therapies unique to each patient's tumor
  - 4 ongoing trials
  - >450 patients treated[2]
  - 18,000 neoantigens selected[2]
- Computational extension of immunotherapy target space[3]
- Semi-automated manufacturing capabilities for iNeST[1]

BIONTECH

AI

## AI-powered bio-engineering

- Development of novel DeepChain platform combining cutting-edge AI and bio-engineering
- Optimization of mRNA design & structure
- Automated dry-wet lab to enhance discovery capabilities
- In-house supercomputing cluster with ~500 PetaFLOPS of Nvidia H100 GPUs

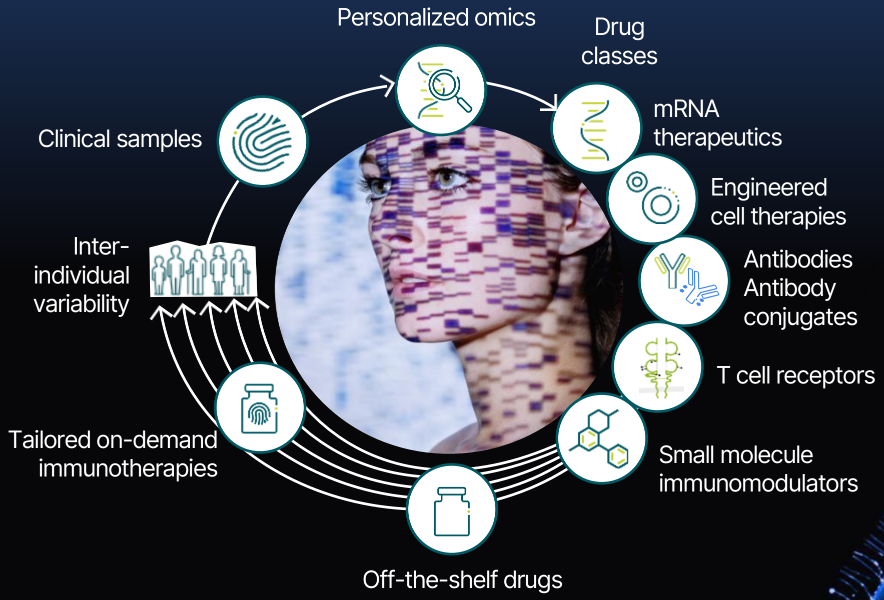
[1] Partnered with Genentech, a member of the Roche Group; [2] From trials BNT122-01, GO39733, GO40558 and ML41081; [3] Castle et al. 2011 *Cancer Res.*

# BioNTech is uniquely positioned with complete AI integration and personalized medicine capabilities under one roof

## Fully-integrated tech-bio company

- Deep genomics & immunology expertise to analyze patient data
- Individualized treatment platforms to address inter-individual variability
- AI-infused & digitally-integrated target & drug discovery and development
- Automated in-house manufacturing to serve patients on time and globally

## Capabilities to build tomorrow's personalized precision medicines



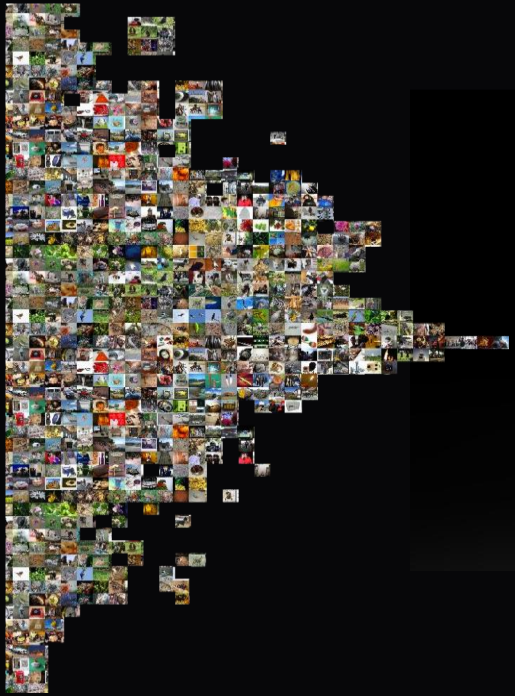
Personalized omics
Drug classes
mRNA therapeutics
Engineered cell therapies
Antibodies Antibody conjugates
T cell receptors
Small molecule immunomodulators
Off-the-shelf drugs
Tailored on-demand immunotherapies
Inter-individual variability
Clinical samples

# Developing the future of AI at BioNTech

Karim Beguir
Co-Founder & CEO
**InstaDeep**

# AI is not a single exponential but a *triple* exponential

## DATA

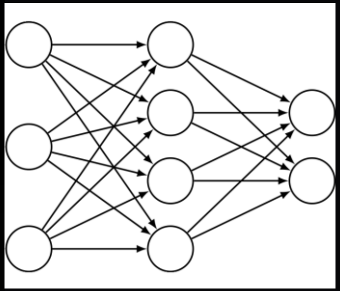


+

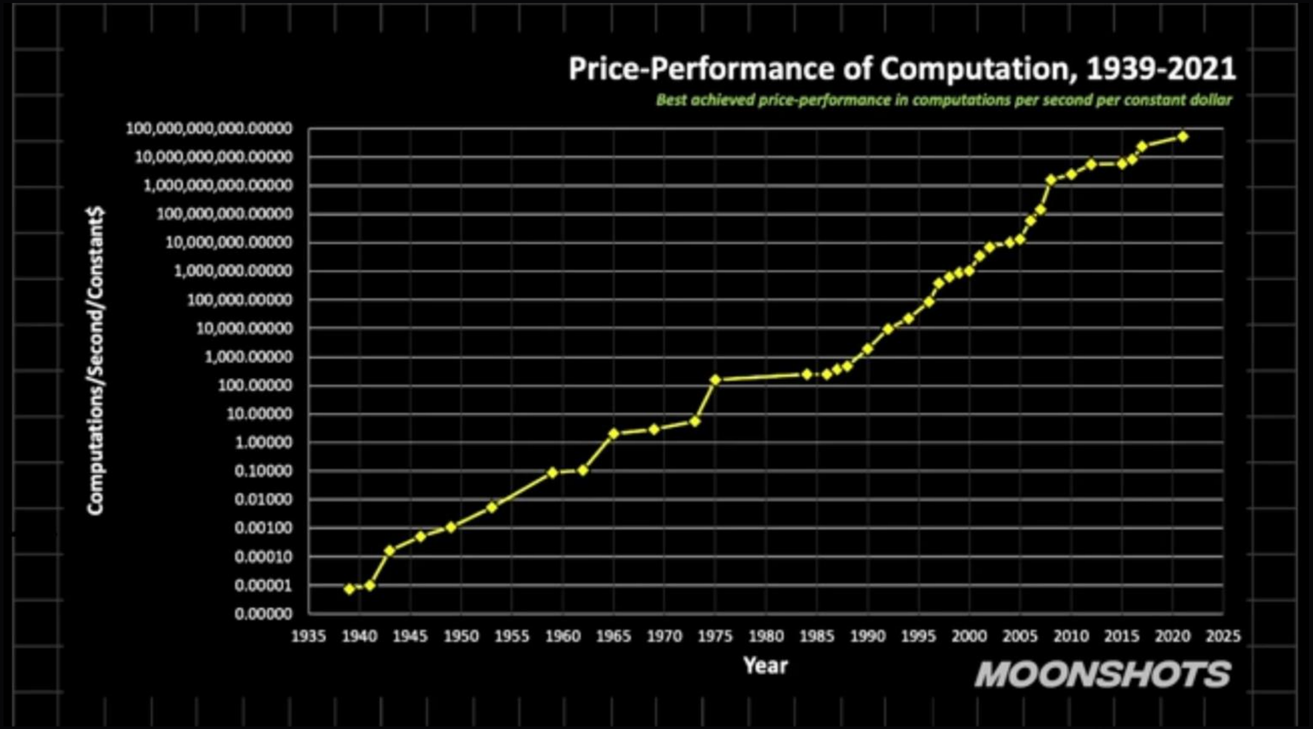
## COMPUTE

+

## MODELS

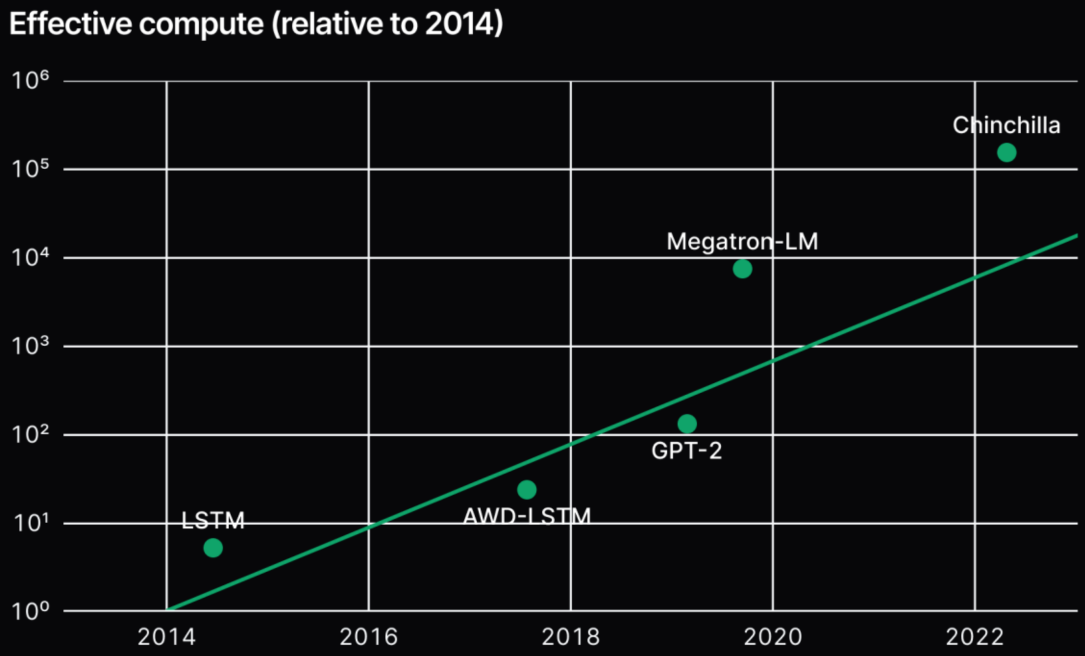


# Moore’s Law: efficiency of hardware compute doubles every two years



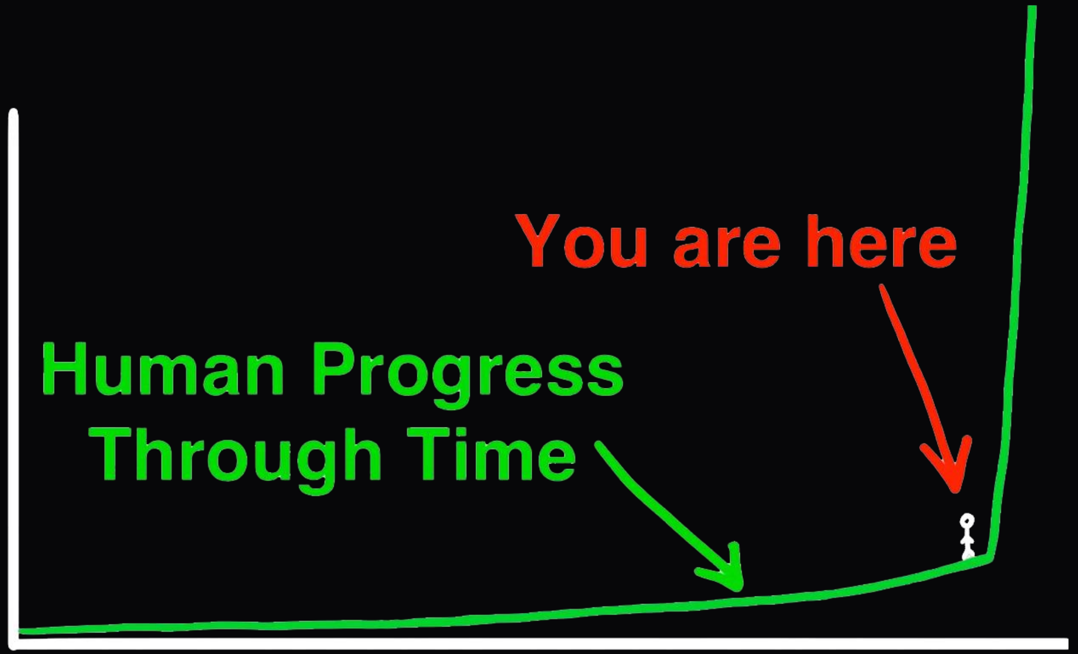
1. Ray Kurzweil Q&A - The Singularity, Human-Machine Integration & AI | EP #83 - Peter Diamandis & Ray Kurzweil Singularity Q&A - (ref) - March 2024

# While the compute efficiency of AI models is also doubling every 8 months



[1] Situational Awareness June The Decade Ahead Leopold Aschenbrenner - (ref) - June 2024

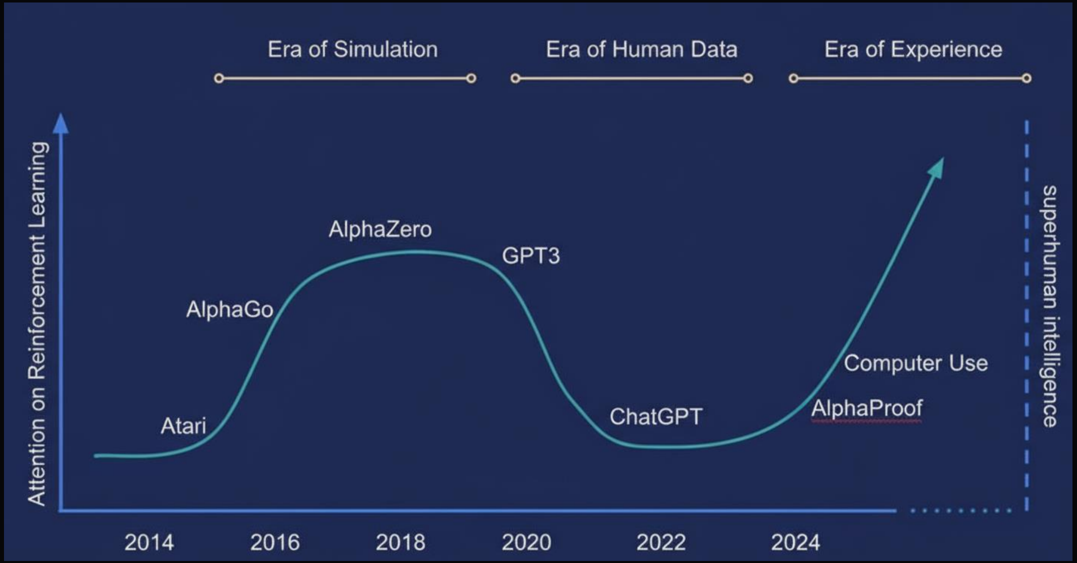
# Is AI likely to expand even further?

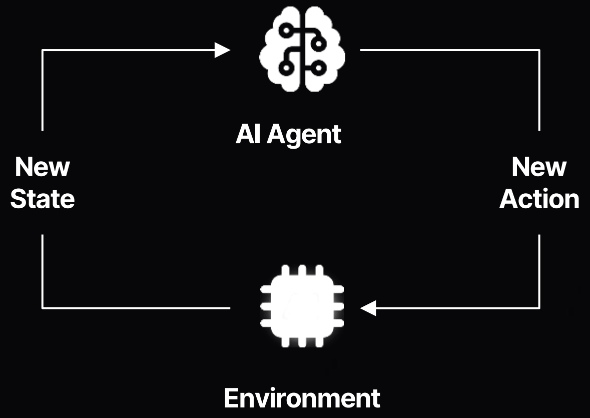


The AI Revolution: The Road to Superintelligence - By Tim Urban -January 22, 2015

# AI *itself* is now accelerating Machine Learning, which creates a new supercycle

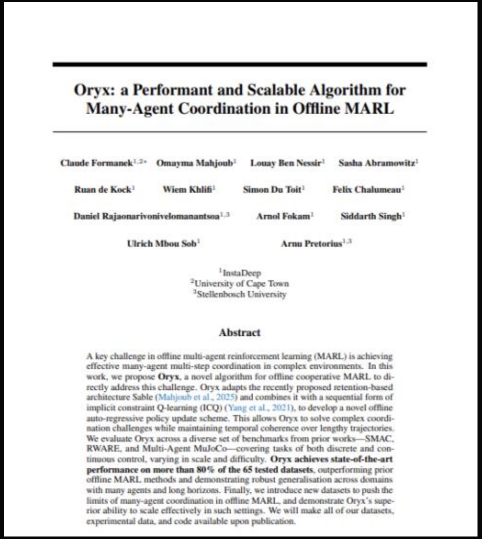
# This supercycle, the Era of Experience, is driven by techniques such as Reinforcement Learning (RL) and Optimization that focus on "learning by doing" with AI agents.





Welcome to the Era of Experience David Silver, Richard S. Sutton - (ref) - Apr 2025

# InstaDeep is active in Reinforcement Learning research



## Oryx: a Performant and Scalable Algorithm for Many-Agent Coordination in Offline MARL

Claude Formanek[1,2*] Omayma Mahjoub[1] Louay Ben Nessir[1] Sasha Abramowitz[1]
Ruan de Kock[1] Wiem Khlifi[1] Simon Du Toit[1] Felix Chalumeau[1]
Daniel Rajaonarivonivelomanantsoa[1,3] Arnol Fokam[1] Siddarth Singh[1]
Ulrich Mbou Sob[1] Arnu Pretorius[1,3]

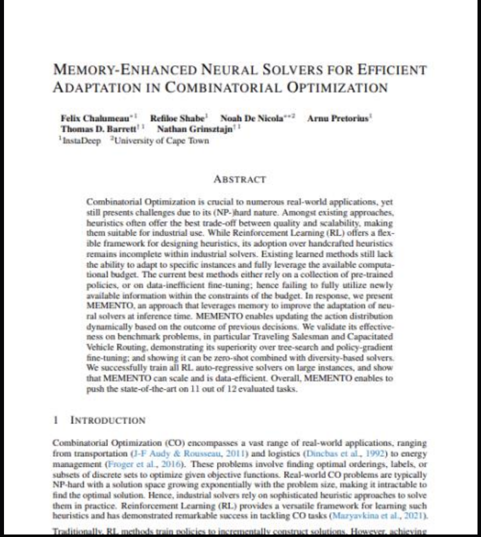
[1]InstaDeep
[2]University of Cape Town
[3]Stellenbosch University

### Abstract

A key challenge in offline multi-agent reinforcement learning (MARL) is achieving effective many-agent multi-step coordination in complex environments. In this work, we propose **Oryx**, a novel algorithm for offline cooperative MARL to directly address this challenge. Oryx adapts the recently proposed retention-based architecture Sable (Mahjoub et al., 2025) and combines it with a sequential form of implicit constraint Q-learning (ICQ) (Yang et al., 2021), to develop a novel offline auto-regressive policy update scheme. This allows Oryx to solve complex coordination challenges while maintaining temporal coherence over lengthy trajectories. We evaluate Oryx across a diverse set of benchmarks from prior works—SMAC, RWARE, and Multi-Agent MuJoCo—covering tasks of both discrete and continuous control, varying in scale and difficulty. **Oryx achieves state-of-the-art performance on more than 80% of the 65 tested datasets**, outperforming prior offline MARL methods and demonstrating robust generalisation across domains with many agents and long horizons. Finally, we introduce new datasets to push the limits of many-agent coordination in offline MARL, and demonstrate Oryx's superior ability to scale effectively in such settings. We will make all of our datasets, experimental data, and code available upon publication.

*Oryx: a Performant and Scalable Algorithm for Many-Agent Coordination in Offline MARL*

***NeurIPS Conference 2025***



## MEMORY-ENHANCED NEURAL SOLVERS FOR EFFICIENT ADAPTATION IN COMBINATORIAL OPTIMIZATION

Felix Chalumeau[*1] Refiloe Shabe[1] Noah De Nicola[**2] Arnu Pretorius[1]
Thomas D. Barrett[†1] Nathan Grinsztajn[†1]
[1]InstaDeep [2]University of Cape Town

### ABSTRACT

Combinatorial Optimization is crucial to numerous real-world applications, yet still presents challenges due to its (NP-)hard nature. Amongst existing approaches, heuristics often offer the best trade-off between quality and scalability, making them suitable for industrial use. While Reinforcement Learning (RL) offers a flexible framework for designing heuristics, its adoption over handcrafted heuristics remains incomplete within industrial solvers. Existing learned methods still lack the ability to adapt to specific instances and fully leverage the available computational budget. The current best methods either rely on a collection of pre-trained policies, or on data-inefficient fine-tuning; hence failing to fully utilize newly available information within the constraints of the budget. In response, we present MEMENTO, an approach that leverages memory to improve the adaptation of neural solvers at inference time. MEMENTO enables updating the action distribution dynamically based on the outcome of previous decisions. We validate its effectiveness on benchmark problems, in particular Traveling Salesman and Capacitated Vehicle Routing, demonstrating its superiority over tree-search and policy-gradient fine-tuning; and showing it can be zero-shot combined with diversity-based solvers. We successfully train all RL auto-regressive solvers on large instances, and show that MEMENTO can scale and is data-efficient. Overall, MEMENTO enables to push the state-of-the-art on 11 out of 12 evaluated tasks.

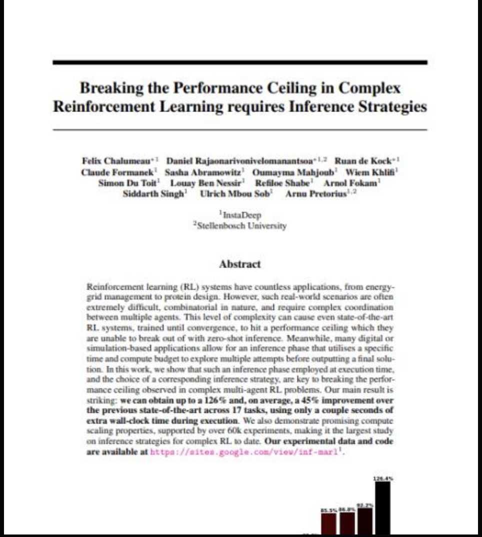
### 1 INTRODUCTION

Combinatorial Optimization (CO) encompasses a vast range of real-world applications, ranging from transportation (J-F Audy & Rousseau, 2011) and logistics (Dincbas et al., 1992) to energy management (Froger et al., 2016). These problems involve finding optimal orderings, labels, or subsets of discrete sets to optimize given objective functions. Real-world CO problems are typically NP-hard with a solution space growing exponentially with the problem size, making it intractable to find the optimal solution. Hence, industrial solvers rely on sophisticated heuristic approaches to solve them in practice. Reinforcement Learning (RL) provides a versatile framework for learning such heuristics and has demonstrated remarkable success in tackling CO tasks (Mazyavkina et al., 2021).

Traditionally, RL methods train policies to incrementally construct solutions. However, achieving

*Memory-Enhanced Neural Solvers for Efficient Adaptation in Combinatorial Optimization*

***NeurIPS Conference 2025 Spotlight***



## Breaking the Performance Ceiling in Complex Reinforcement Learning requires Inference Strategies

Felix Chalumeau[*1] Daniel Rajaonarivonivelomanantsoa[*1,2] Ruan de Kock[*1]
Claude Formanek[1] Sasha Abramowitz[1] Oumayma Mahjoub[1] Wiem Khlifi[1]
Simon Du Toit[1] Louay Ben Nessir[1] Refiloe Shabe[1] Arnol Fokam[1]
Siddarth Singh[1] Ulrich Mbou Sob[1] Arnu Pretorius[1,2]

[1]InstaDeep
[2]Stellenbosch University

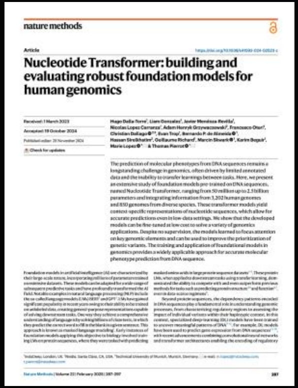
### Abstract

Reinforcement learning (RL) systems have countless applications, from energy-grid management to protein design. However, such real-world scenarios are often extremely difficult, combinatorial in nature, and require complex coordination between multiple agents. This level of complexity can cause even state-of-the-art RL systems, trained until convergence, to hit a performance ceiling which they are unable to break out of with zero-shot inference. Meanwhile, many digital or simulation-based applications allow for an inference phase that utilises a specific time and compute budget to explore multiple attempts before outputting a final solution. In this work, we show that such an inference phase employed at execution time, and the choice of a corresponding inference strategy, are key to breaking the performance ceiling observed in complex multi-agent RL problems. Our main result is striking: **we can obtain up to a 126% and, on average, a 45% improvement over the previous state-of-the-art across 17 tasks, using only a couple seconds of extra wall-clock time during execution**. We also demonstrate promising compute scaling properties, supported by over 60k experiments, making it the largest study on inference strategies for complex RL to date. **Our experimental data and code are available at** https://sites.google.com/view/inf-marl[1].

*Breaking the Performance Ceiling in Complex Reinforcement Learning requires Inference Strategies*

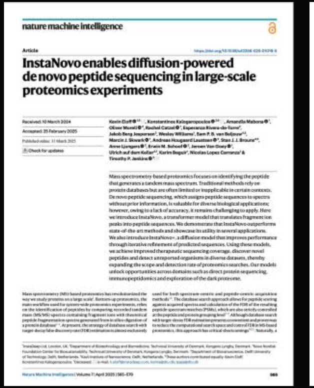
***NeurIPS Conference 2025 Oral***

# Biology and AI know-how: 6 Nature journal publications in less than 12 months


nature methods

Article

Nucleotide Transformer: building and evaluating robust foundation models for human genomics


nature machine intelligence

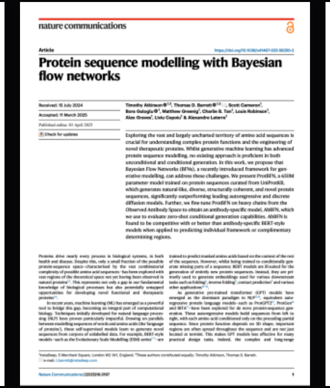
Article

InstaNovo enables diffusion-powered de novo peptide sequencing in large-scale proteomics experiments

nature machine intelligence

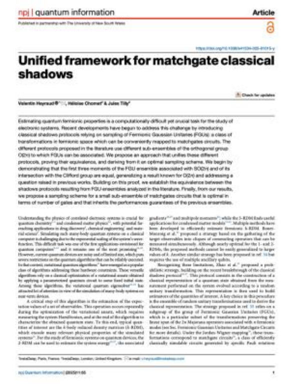
Article

A multimodal conversational agent for DNA, RNA and protein tasks


nature communications

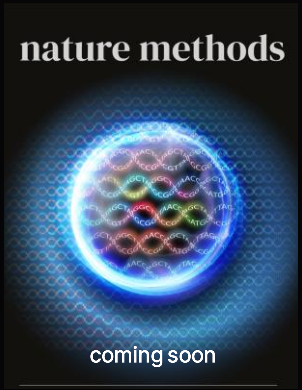
Article

Protein sequence modelling with Bayesian flow networks


npj | quantum information

Article

Unified framework for matchgate classical shadows


nature methods

coming soon

***Nucleotide Transformer***

Building and Evaluating Robust Foundation Models for Human Genomics

***Nature Methods***
***2024***

***InstaNovo***

ML for *de novo* peptide sequencing for large-scale mass spectrometry proteomics

***Nature Machine Intelligence***
***2025***

***ChatNT***

A Multi-Modal Conversational Agent for Genomics

***Nature Machine Intelligence***
***2025***

***ProtBFN & AbBFN***

Protein Sequence Modelling with Bayesian Flow Networks

***Nature Communications***
***2025***

***Matchgate Classical Shadows***

Unified Matchgate Classical Shadows for Quantum Fermionic Systems

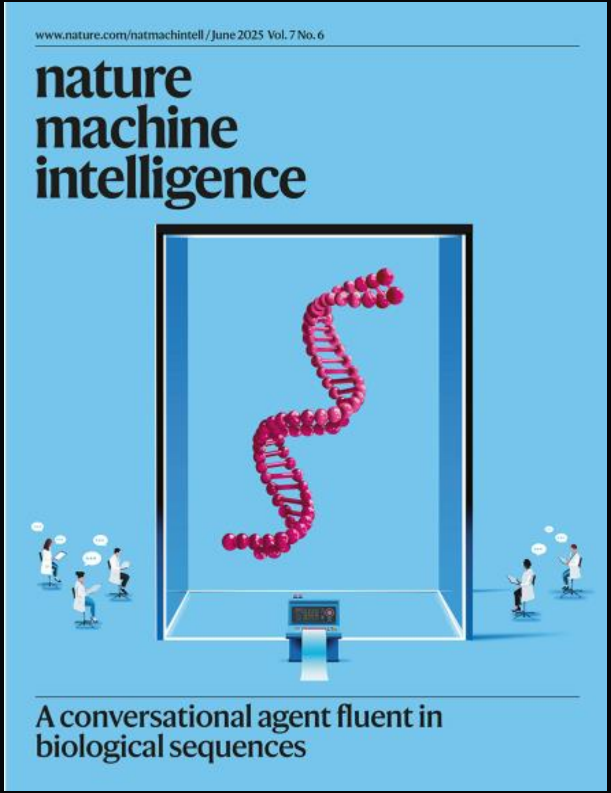
***Nature Partner Journals Quantum Information***
***2025***

***SegmentNT***

Annotating the genome at single-nucleotide resolution with DNA foundation model

***Nature Methods***
***2025***

# ChatNT, our conversational agent for Genomics, made the cover of Nature Machine Intelligence



www.nature.com/natmachintell / June 2025 Vol. 7 No. 6

nature
machine
intelligence

A conversational agent fluent in
biological sequences

# InstaDeep and BioNTech are building across the full stack of AI:

Compute & Model Scaling

AI Innovation

Data Acquisition & Refinement

Applications

# Compute & Model Scaling

Alex Laterre
Head of AI Research
InstaDeep

# Scaling laws drive AI innovation

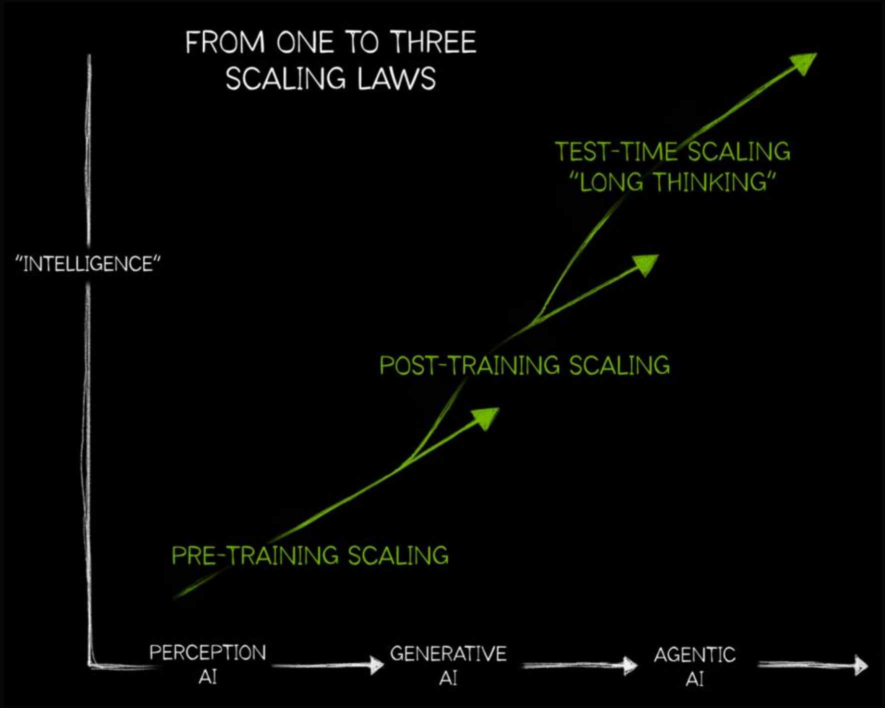
### Gold medal at IMO 2025 achieved[1]

### 1st place at International Programming Contest[2]

### Growing capabilities for physical agents[3]

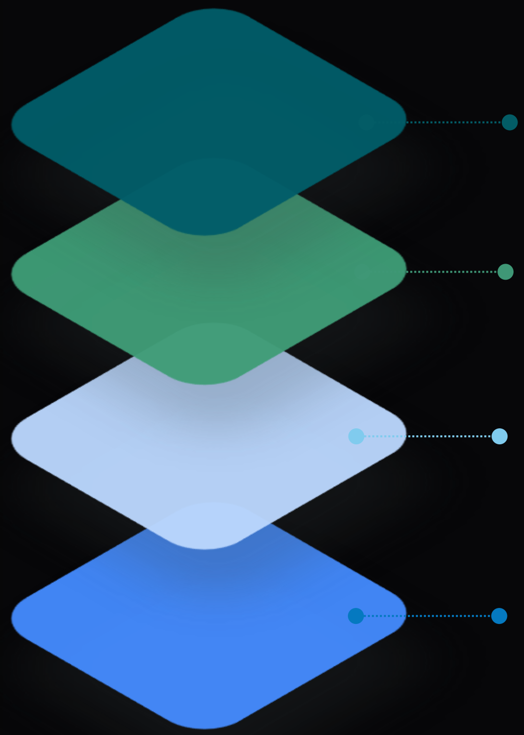
1. DeepMind achieves gold medal-level performance on the 2025 International Mathematical Olympiad with a general-purpose reasoning LLM! (ref) – 21st of July 2025
2. OpenAI general-purpose reasoning models solved all 12 problems at the 2025 International Collegiate Programming Contest (ICPC) World Finals (ref) – 17th of September 2025
3. DeepMind released Gemini Robotics 1.5 - AI models that let robots perceive, plan, think and act across diverse physical environments to complete complex, multi-step tasks with explainable reasoning (ref ) – 25th of September 2025

# Compounding intelligence



Jensen Keynote that Nvidia GTC 2025

# A fully integrated AI ecosystem



## AI innovation

*Pioneer work in generative models, representation learning, and reasoning.*

## ML software ecosystem in JAX

*Software for high-performance computing and advanced model optimization.*

## AIChor orchestration platform

*A Kubernetes-native AI training platform for seamless scaling and fast experimentation.*

## InstaDeep's AI supercomputer

*Purpose-built for AI, delivering full control, visibility, and performance.*

# Kyber, InstaDeep's AI supercomputer

**~500 PetaFLOPS**
of Nvidia H100 GPUs

**86,000**
CPU Cores

**1.2 Tons**
of Hardware per Rack

- Custom rack design engineered in-house
- **Optimised for AI performance** and cost efficiency
- Powered 100% by renewable energy
- Designed to scale seamlessly with next-generation hardware
- Tight **hardware–software integration** for control and efficiency

# AIchor orchestration platform

**AIchor, a complete AI training platform, ready for production and built for scale.**

### Simple
GitOps workflow: Commit → Build → Run → Monitor

### Scalable
Kubernetes-native provisioning and auto-scaling across clusters

### Flexible
Modular data plane for multi-cluster and multi-cloud compute

https://aichor.ai/

**+15,000**
experiments / month in 2025

**+75%**
GPU usage

# ML software ecosystem in JAX

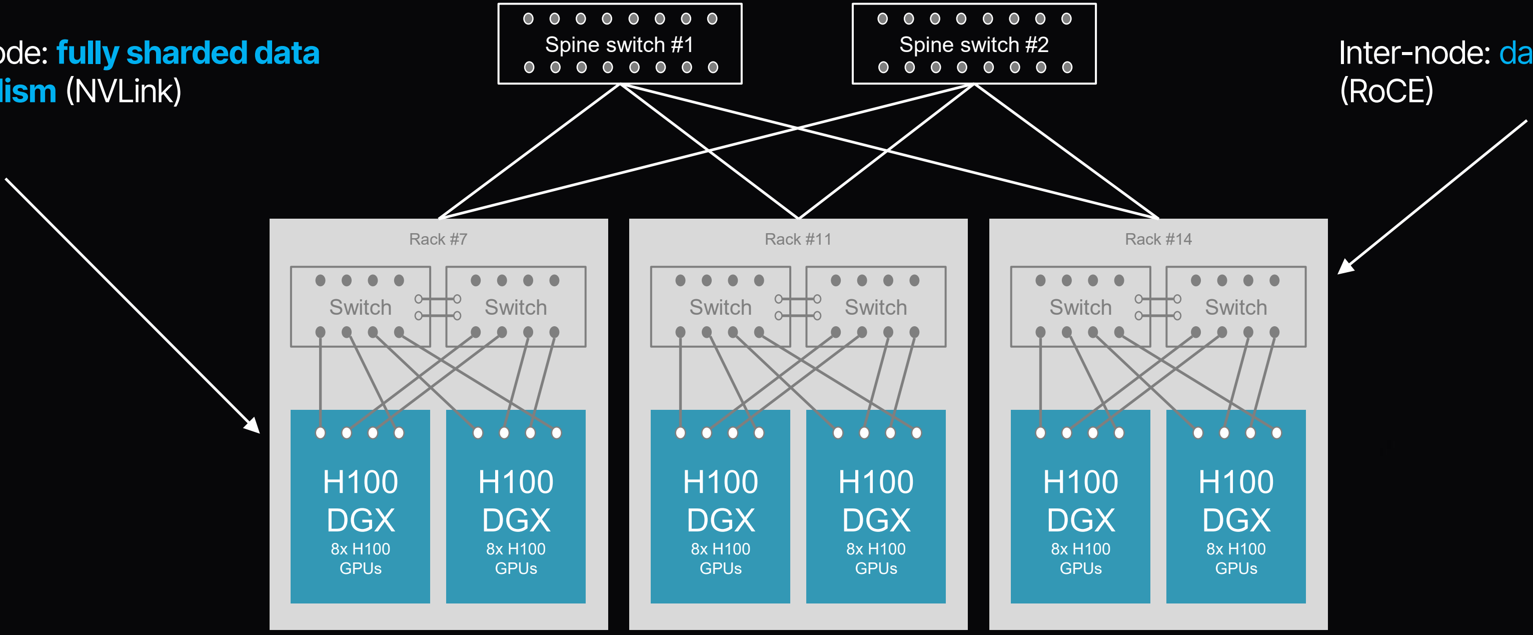
- Scale → from rapid prototyping to large-scale training and deployment
- Efficiency → "better, faster, cheaper" AI workloads that maximise hardware usage
- Modularity → interoperable, reliable, and optimised components working together

# 1 | Efficiently train 100B-parameter foundation models

## Hierarchical model sharding

Intra-node: **fully sharded data parallelism** (NVLink)

Inter-node: data parallelism (RoCE)



Kyber

# 1 | Efficiently train 100B-parameter foundation models

## Hierarchical model sharding

- ✓ Intra-node: fully sharded data parallelism (NVLink)
- ✓ Inter-node: data parallelism (RoCE)
- ✓ *Tensor and sequence parallelism available*

## Code optimizations

- ✓ CuDNN kernels (e.g. Flash Attention)
- ✓ Mixed precision with FP8 quantisation
- ✓ XLA compiler and RoCE configuration tuning
- ✓ NUMA binding affinity
- ✓ …

**+66 %**

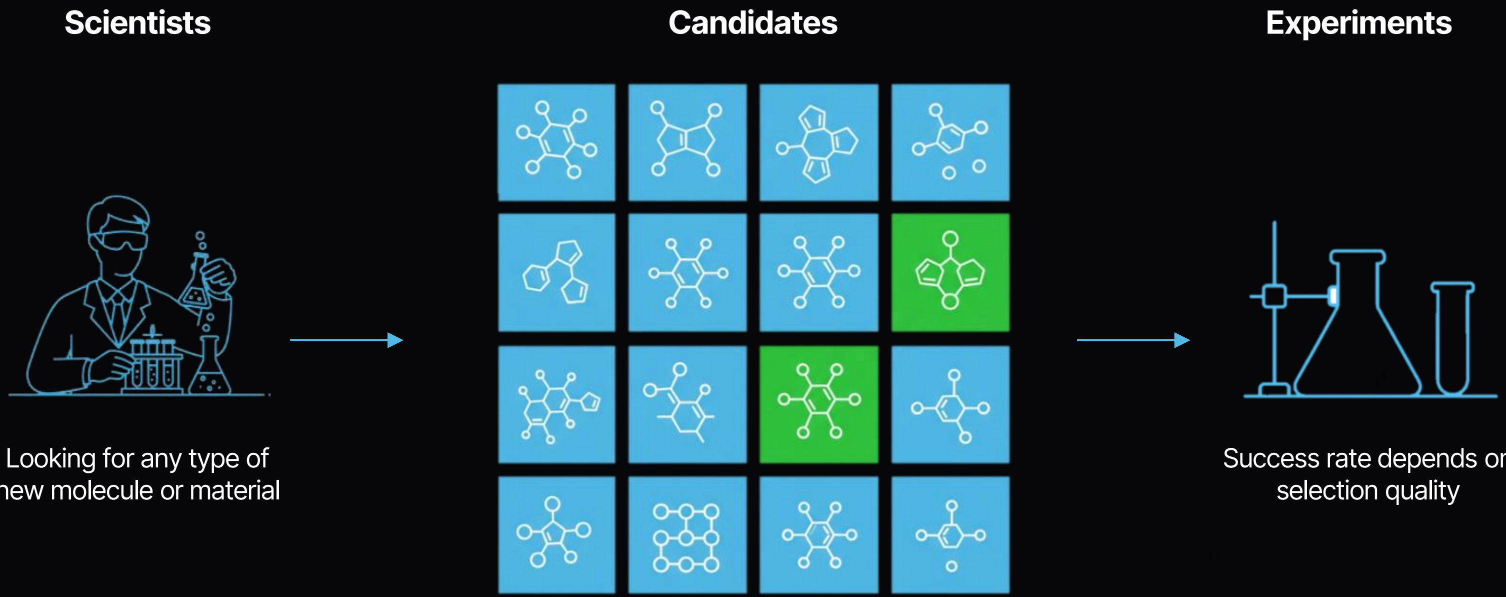
Model FLOPs Utilization (MFU) on 64 x H100 GPUs

**Model FLOPs Utilization (MFU)**
This is the ratio of observed throughput (tokens per second) to the theoretical maximum throughput of a system running at peak FLOPs.

e.g, Llama 3.1 405B achieves 38 to 41% MFU on 16,384 H100 GPUs.

# 2 | Scaling molecule screening with Machine Learning Interatomic Potential

Simulating molecular properties at scale is key to many industries, including drug discovery, materials, chemicals. Machine Learning Interatomic Potentials allow **quantum accuracy orders of magnitude faster** on molecular simulations



**Scientists**

Looking for any type of new molecule or material

**Candidates**

**Experiments**

Success rate depends on selection quality

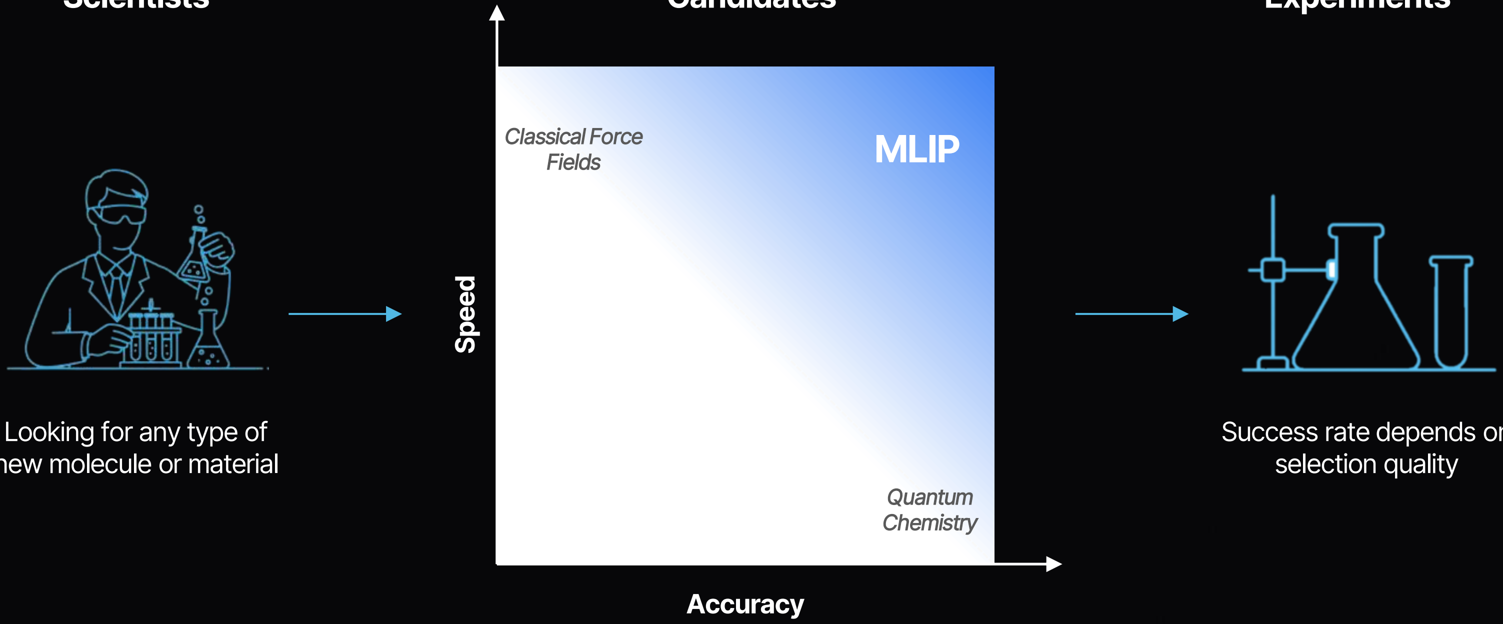
# 2 | Scaling molecule screening with Machine Learning Interatomic Potential

Simulating molecular properties at scale is key to many industries, including drug discovery, materials, chemicals. Machine Learning Interatomic Potentials allow **quantum accuracy orders of magnitude faster** on molecular simulations

**Scientists**

**Candidates**

**Experiments**



Looking for any type of new molecule or material

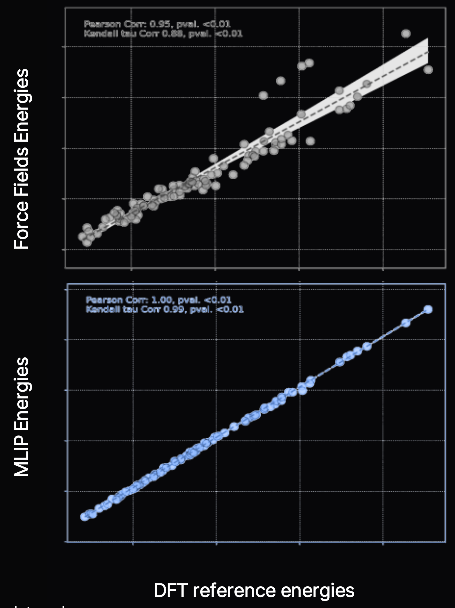
Success rate depends on selection quality

Internal

# 2 | Scaling molecule screening with Machine Learning Interatomic Potential

## Better

Quantum Chemistry-level accuracy



## Cheaper

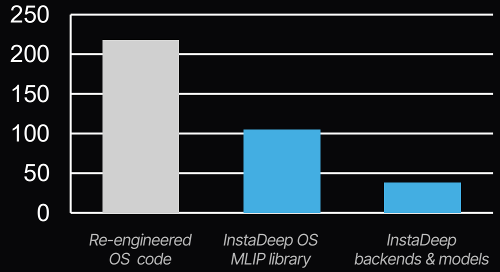
+10,000 times cheaper than DFT

*Estimated on a 150 atoms molecule*

| Method | Hardware | Runtime | Relative Compute cost |
|---|---|---|---|
| DFT | 64 CPU cores | ~ 145 days | **$12,500** |
| MLIP | 1 H100 GPU | < 20 min | **$1** |

Up to 5x speed-up in simulation speed

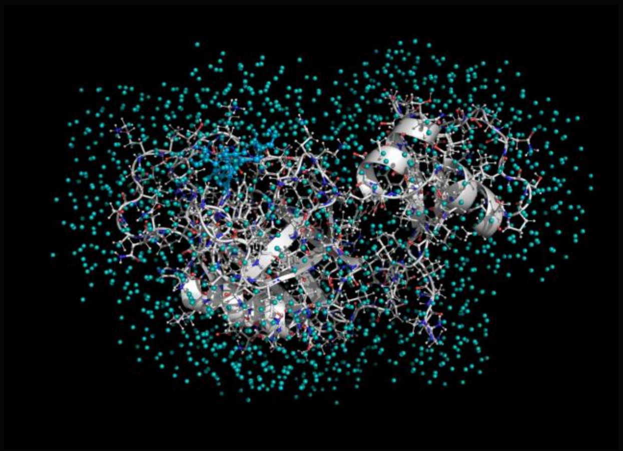
*160 atoms molecule for 1ns (runtime in min)*



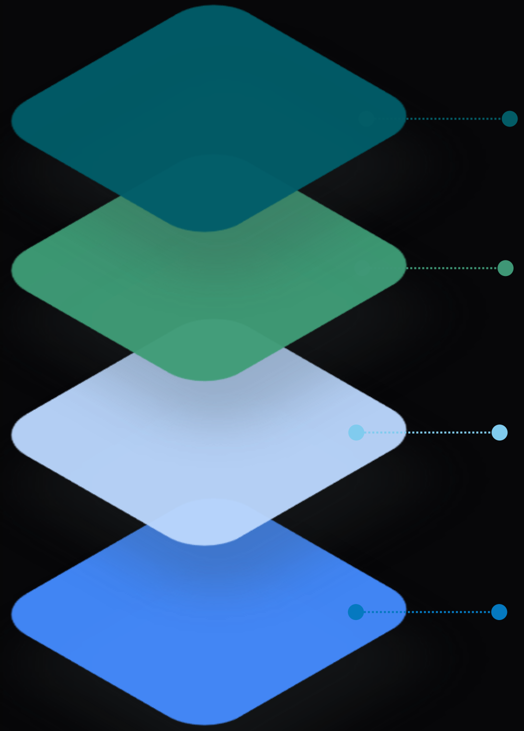
## Scalable

Over 100,000 atoms on one GPU

*Imipenem binding to L,D-transpeptidase*



# A fully integrated AI ecosystem



## AI innovation

*Pioneer work in generative models, representation learning, and reasoning.*

## ML software ecosystem in JAX

*Software for high-performance computing and advanced model optimization.*

## AIChor orchestration platform

*A Kubernetes-native AI training platform for seamless scaling and fast experimentation.*

## InstaDeep's AI supercomputer

*Purpose-built for AI, delivering full control, visibility, and performance.*

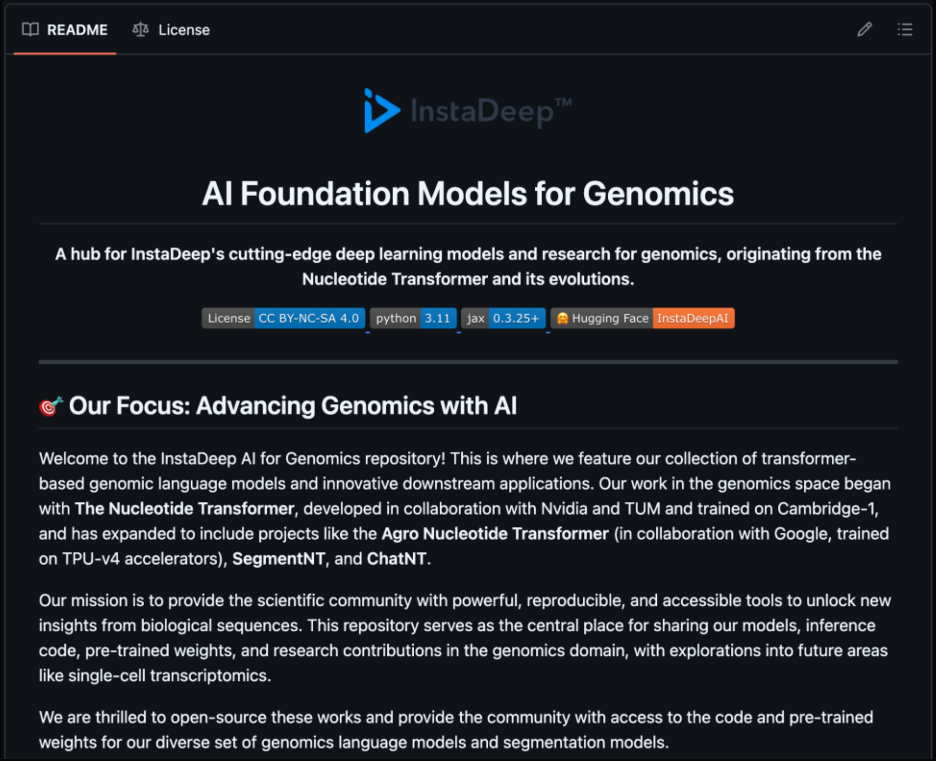
# AI Innovation

AI innovation

# Generative AI for genomics

Bernardo Almeida
Senior Research Scientist
InstaDeep

# Nucleotide Transformer: one of the most popular genomics AI models on Hugging Face



## AI Foundation Models for Genomics

**A hub for InstaDeep's cutting-edge deep learning models and research for genomics, originating from the Nucleotide Transformer and its evolutions.**

### 🎯 Our Focus: Advancing Genomics with AI

Welcome to the InstaDeep AI for Genomics repository! This is where we feature our collection of transformer-based genomic language models and innovative downstream applications. Our work in the genomics space began with **The Nucleotide Transformer**, developed in collaboration with Nvidia and TUM and trained on Cambridge-1, and has expanded to include projects like the **Agro Nucleotide Transformer** (in collaboration with Google, trained on TPU-v4 accelerators), **SegmentNT**, and **ChatNT**.

Our mission is to provide the scientific community with powerful, reproducible, and accessible tools to unlock new insights from biological sequences. This repository serves as the central place for sharing our models, inference code, pre-trained weights, and research contributions in the genomics domain, with explorations into future areas like single-cell transcriptomics.

We are thrilled to open-source these works and provide the community with access to the code and pre-trained weights for our diverse set of genomics language models and segmentation models.

## +1 Million Downloads

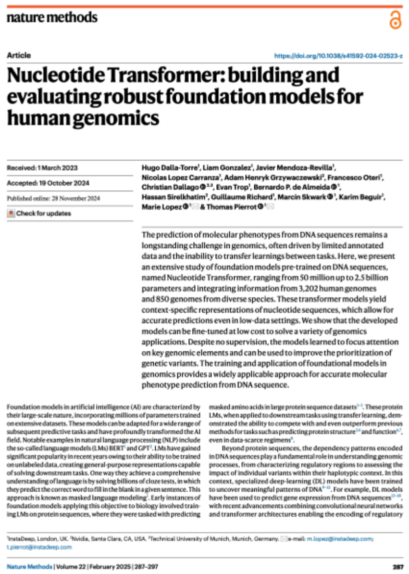
Across model sizes[1]

## +500 Citations

Across model types[2]

[1] Hugging Face Statistics.
[2] Google Scholar.

# Generative AI for genomics – publications in top-tier journals



nature methods

Article https://doi.org/10.1038/s41592-024-02523-z

## Nucleotide Transformer: building and evaluating robust foundation models for human genomics

Received: 1 March 2023

Accepted: 19 October 2024

Published online: 28 November 2024

Check for updates

Hugo Dalla-Torre[1], Liam Gonzalez[1], Javier Mendoza-Revilla[1], Nicolas Lopez Carranza[1], Adam Henryk Grzywaczewski[2], Francesco Oteri[1], Christian Dallago[2,3], Evan Trop[1], Bernardo P. de Almeida[1], Hassan Sirelkhatim[2], Guillaume Richard[1], Marcin Skwark[1], Karim Beguir[1], Marie Lopez[1] & Thomas Pierrot[1]

The prediction of molecular phenotypes from DNA sequences remains a longstanding challenge in genomics, often driven by limited annotated data and the inability to transfer learnings between tasks. Here, we present an extensive study of foundation models pre-trained on DNA sequences, named Nucleotide Transformer, ranging from 50 million up to 2.5 billion parameters and integrating information from 3,202 human genomes and 850 genomes from diverse species. These transformer models yield context-specific representations of nucleotide sequences, which allow for accurate predictions even in low-data settings. We show that the developed models can be fine-tuned at low cost to solve a variety of genomics applications. Despite no supervision, the models learned to focus attention on key genomic elements and can be used to improve the prioritization of genetic variants. The training and application of foundational models in genomics provides a widely applicable approach for accurate molecular phenotype prediction from DNA sequence.

Foundation models in artificial intelligence (AI) are characterized by their large-scale nature, incorporating millions of parameters trained on extensive datasets. These models can be adapted for a wide range of subsequent predictive tasks and have profoundly transformed the AI field. Notable examples in natural language processing (NLP) include the so-called language models (LMs) BERT[1] and GPT[2]. LMs have gained significant popularity in recent years owing to their ability to be trained on unlabeled data, creating general-purpose representations capable of solving downstream tasks. One way they achieve a comprehensive understanding of language is by solving billions of cloze tests, in which they predict the correct word to fill in the blank in a given sentence. This approach is known as masked language modeling[1]. Early instances of foundation models applying this objective to biology involved training LMs on protein sequences, where they were tasked with predicting masked amino acids in large protein sequence datasets[3–5]. These protein LMs, when applied to downstream tasks using transfer learning, demonstrated the ability to compete with and even outperform previous methods for tasks such as predicting protein structure[3,4] and function[6,7], even in data-scarce regimens[8].

Beyond protein sequences, the dependency patterns encoded in DNA sequences play a fundamental role in understanding genomic processes, from characterizing regulatory regions to assessing the impact of individual variants within their haplotypic context. In this context, specialized deep-learning (DL) models have been trained to uncover meaningful patterns of DNA[9–12]. For example, DL models have been used to predict gene expression from DNA sequences[13–18], with recent advancements combining convolutional neural networks and transformer architectures enabling the encoding of regulatory

[1]InstaDeep, London, UK. [2]Nvidia, Santa Clara, CA, USA. [3]Technical University of Munich, Munich, Germany. e-mail: m.lopez@instadeep.com; t.pierrot@instadeep.com

Nature Methods | Volume 22 | February 2025 | 287–297 287

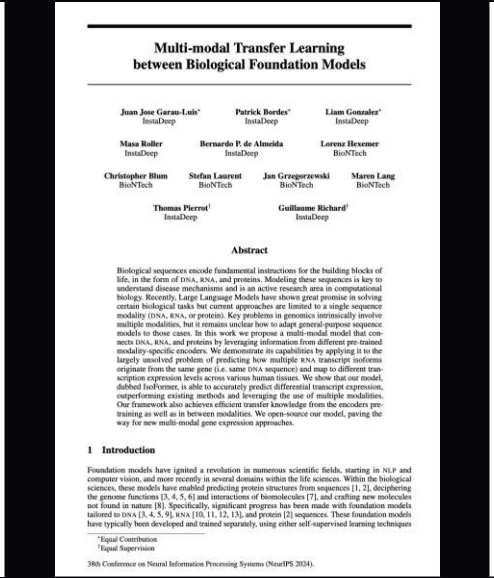
## Annotating the genome at single-nucleotide resolution with DNA foundation models

Bernardo P. de Almeida[*,1], Hugo Dalla-Torre[*,1], Guillaume Richard[1], Christopher Blum[2], Lorenz Hexemer[2], Maxence Gélard[1], Javier Mendoza-Revilla[1], Ziqi Tang[1,3], Frederikke I. Marin[1,4], David M. Emms[1], Priyanka Pandey[2], Stefan Laurent[2], Marie Lopez[1], Alexandre Laterre[1], Maren Lang[2], Uğur Şahin[2], Karim Beguir[1] and Thomas Pierrot[1]

[*]Equal contributions, [1]InstaDeep, [2]BioNTech, [3]Cold Spring Harbor Laboratory, [4]University of Copenhagen

Genome annotation models that directly analyze DNA sequences are indispensable for modern biological research, enabling rapid and accurate identification of genes and other functional elements. This capability is paramount as the volume of sequenced genomes rapidly expands, making the need for efficient and accurate annotation methods increasingly critical, particularly in the context of genetic variant prediction and in-silico sequence design. Current annotation tools are typically developed for specific element classes and trained from scratch using supervised learning on datasets that are often limited in size. This approach constrains their performance and ability to generalize to new genomes. Here, we frame the genome annotation problem as instance segmentation and introduce a novel methodology for fine-tuning pre-trained DNA foundation models to segment 14 different genic and regulatory elements at single-nucleotide resolution. We leverage the self-supervised pre-trained model Nucleotide Transformer (NT) to develop a general segmentation model, SegmentNT, capable of processing DNA sequences up to 50kb long. By utilizing pre-trained weights from NT, SegmentNT surpasses the performance of several ablation models and baselines, including convolutional networks with one-hot encoded nucleotide sequences and large models trained from scratch. We demonstrate state-of-the-art performance on gene annotation, splice site and regulatory elements detection throughout the genome. We also leveraged our framework to accommodate two extra DNA foundation models, Enformer and Borzoi, extending the sequence context up to 500kb and enhancing performance on regulatory elements. Finally, we show that a SegmentNT model trained on human genomic elements generalizes to elements of different species, and a multi-species SegmentNT model achieves strong generalization across unseen species. Our approach is readily extensible to additional genomic elements and species. We have made our SegmentNT human and multi-species models, as well as the SegmentEnformer and SegmentBorzoi models, available on our github repository in Jax and HuggingFace space in Pytorch.

Corresponding author(s): t.pierrot@instadeep.com



## Multi-modal Transfer Learning between Biological Foundation Models

Juan Jose Garau-Luis[*] InstaDeep
Patrick Bordes[*] InstaDeep
Liam Gonzalez[*] InstaDeep
Masa Roller InstaDeep
Bernardo P. de Almeida InstaDeep
Lorenz Hexemer BioNTech
Christopher Blum BioNTech
Stefan Laurent BioNTech
Jan Grzegorzewski BioNTech
Maren Lang BioNTech
Thomas Pierrot[†] InstaDeep
Guillaume Richard[†] InstaDeep

### Abstract

Biological sequences encode fundamental instructions for the building blocks of life, in the form of DNA, RNA, and proteins. Modeling these sequences is key to understand disease mechanisms and is an active research area in computational biology. Recently, Large Language Models have shown great promise in solving certain biological tasks but current approaches are limited to a single sequence modality (DNA, RNA, or protein). Key problems in genomics intrinsically involve multiple modalities, but it remains unclear how to adapt general-purpose sequence models to those cases. In this work we propose a multi-modal model that connects DNA, RNA, and proteins by leveraging information from different pre-trained modality-specific encoders. We demonstrate its capabilities by applying it to the largely unsolved problem of predicting how multiple RNA transcript isoforms originate from the same gene (i.e. same DNA sequence) and map to different transcription expression levels across various human tissues. We show that our model, dubbed IsoFormer, is able to accurately predict differential transcript expression, outperforming existing methods and leveraging the use of multiple modalities. Our framework also achieves efficient transfer knowledge from the encoders pre-training as well as in between modalities. We open-source our model, paving the way for new multi-modal gene expression approaches.

### 1 Introduction

Foundation models have ignited a revolution in numerous scientific fields, starting in NLP and computer vision, and more recently in several domains within the life sciences. Within the biological sciences, these models have enabled predicting protein structures from sequences [1, 2], deciphering the genome functions [3, 4, 5, 6] and interactions of biomolecules [7], and crafting new molecules not found in nature [8]. Specifically, significant progress has been made with foundation models tailored to DNA [3, 4, 5, 9], RNA [10, 11, 12, 13], and protein [2] sequences. These foundation models have typically been developed and trained separately, using either self-supervised learning techniques

[*]Equal Contribution
[†]Equal Supervision

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

nature machine intelligence

Article https://doi.org/10.1038/s42256-025-01047-1

## A multimodal conversational agent for DNA, RNA and protein tasks

Received: 15 May 2024

Accepted: 30 April 2025

Published online: 6 June 2025

Check for updates

Bernardo P. de Almeida[1,4], Guillaume Richard[1,4], Hugo Dalla-Torre[1], Christopher Blum[2], Lorenz Hexemer[2], Priyanka Pandey[2], Stefan Laurent[2], Chandana Rajesh[3], Marie Lopez[1], Alexandre Laterre[1], Maren Lang[2], Uğur Şahin[2], Karim Beguir[1] & Thomas Pierrot[1]

Language models are thriving, powering conversational agents that assist and empower humans to solve a number of tasks. Recently, these models were extended to support additional modalities including vision, audio and video, demonstrating impressive capabilities across multiple domains, including healthcare. Still, conversational agents remain limited in biology as they cannot yet fully comprehend biological sequences. Meanwhile, high-performance foundation models for biological sequences have been built through self-supervision over sequencing data, but these need to be fine-tuned for each specific application, preventing generalization between tasks. In addition, these models are not conversational, which limits their utility to users with coding capabilities. Here we propose to bridge the gap between biology foundation models and conversational agents by introducing ChatNT, a multimodal conversational agent with an advanced understanding of biological sequences. ChatNT achieves new state-of-the-art results on the Nucleotide Transformer benchmark while being able to solve all tasks at once, in English, and to generalize to unseen questions. In addition, we have curated a set of more biologically relevant instruction tasks from DNA, RNA and proteins, spanning multiple species, tissues and biological processes. ChatNT reaches performance on par with state-of-the-art specialized methods on those tasks. We also present a perplexity-based technique to help calibrate the confidence of our model predictions. By applying attribution methods through the English decoder and DNA encoder, we demonstrate that ChatNT's answers are based on biologically coherent features such as detecting the promoter TATA motif or splice site dinucleotides. Our framework for genomics instruction tuning can be extended to more tasks and data modalities (for example, structure and imaging), making it a widely applicable tool for biology. ChatNT provides a potential direction for building generally capable agents that understand biology from first principles while being accessible to users with no coding background.

[1]InstaDeep, London, UK. [2]BioNTech, Mainz, Germany. [3]CSHL, New York, NY, USA. [4]These authors contributed equally: Bernardo P. de Almeida, Guillaume Richard. e-mail: t.pierrot@instadeep.com

Nature Machine Intelligence | Volume 7 | June 2025 | 928–941 928

## Nucleotide Transformer

Building and Evaluating Robust **Foundation Models** for Human Genomics

*Nature Methods, 2024*

## SegmentNT

Segmenting the Genome at **Single-Nucleotide Resolution** with DNA Foundational Models

***Nature Methods, 2025***

## Isoformer

**Multi-modal** Transfer Learning between Biological Foundation Models

*NeurIPS Conference, 2024*

## ChatNT

A Multi-Modal Conversational Agent for Genomics

***Nature Machine Intelligence, 2025***

# Exploiting the data available with the aim of building a best-in-class model for genomics

NT
Evo

Learn from genomes

OR

Borzoi
AlphaGenome

Learn from functional data

**Nucleotide Transformer v3**
NTv3

Pre-training on genomes from >150,000 species

Post-training on >17,000 functional tracks across 16 species

# Introducing NTv3: a new, truly foundational, model for genomics with a million nucleotide context

**Multi-species**
more than 150,000 species genomes

**Multimodal**
genomes + functional tracks + genome annotation

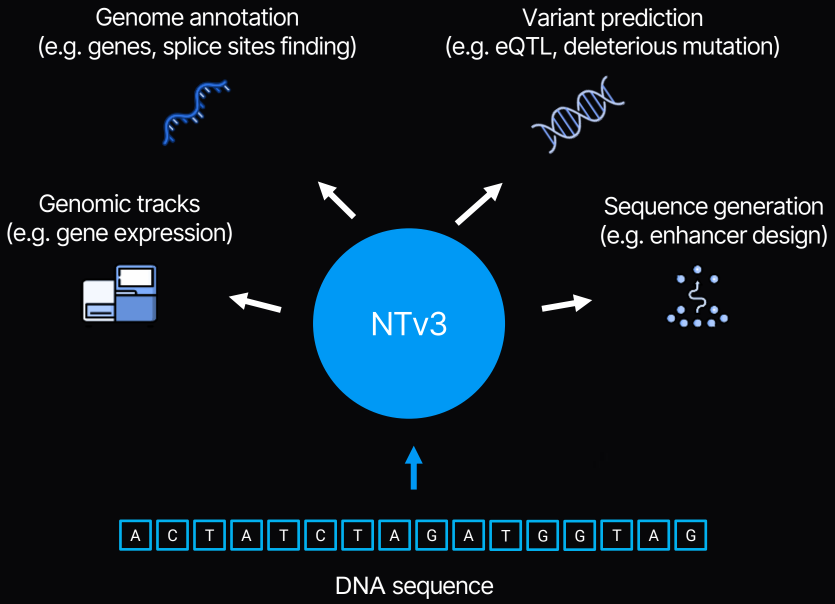
**Multi-domains**
human genomics, plants genomics, metagenomics

**Long-range**
up to 1 million input nucleotides

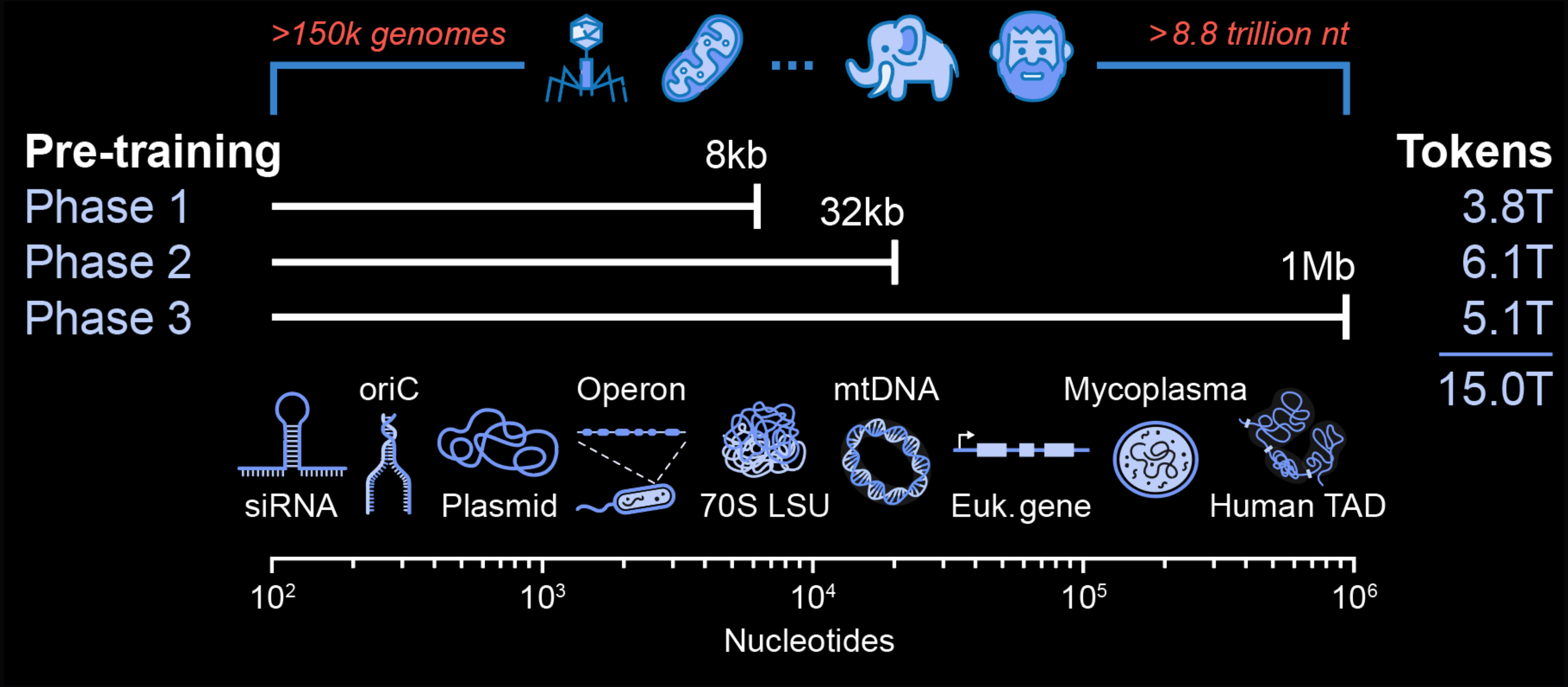
**Generative capacities**
design of DNA sequences *de novo* with **in-vitro validation**

**Suite of models**
from 10M → 4B parameters

**Designed for efficiency**
fastest foundation models available

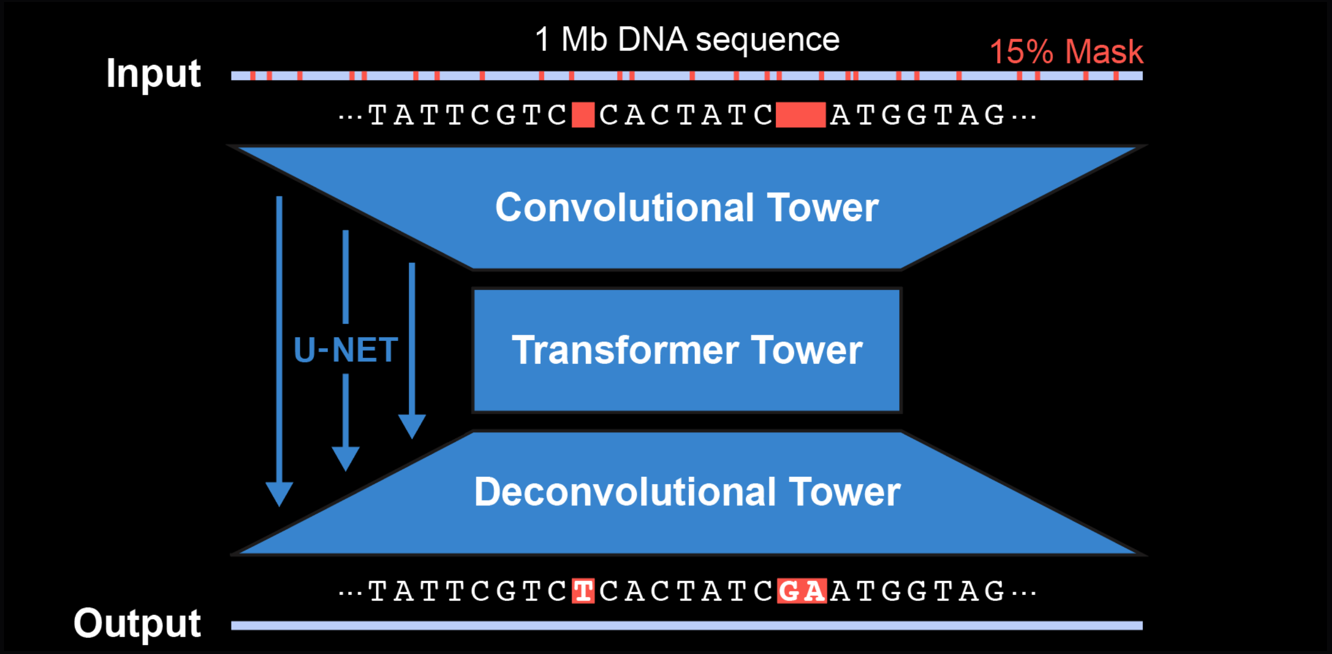


# Pre-training | Learning from +150,000 species genomes

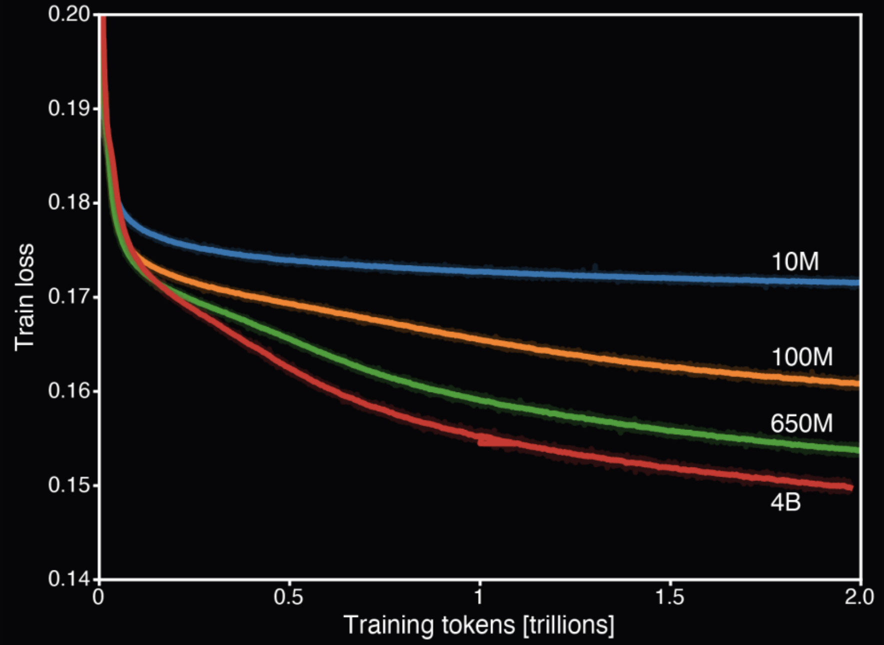


# Pre-training | Learning from +150,000 species genomes

## NTv3 learns through Masked Language Modelling

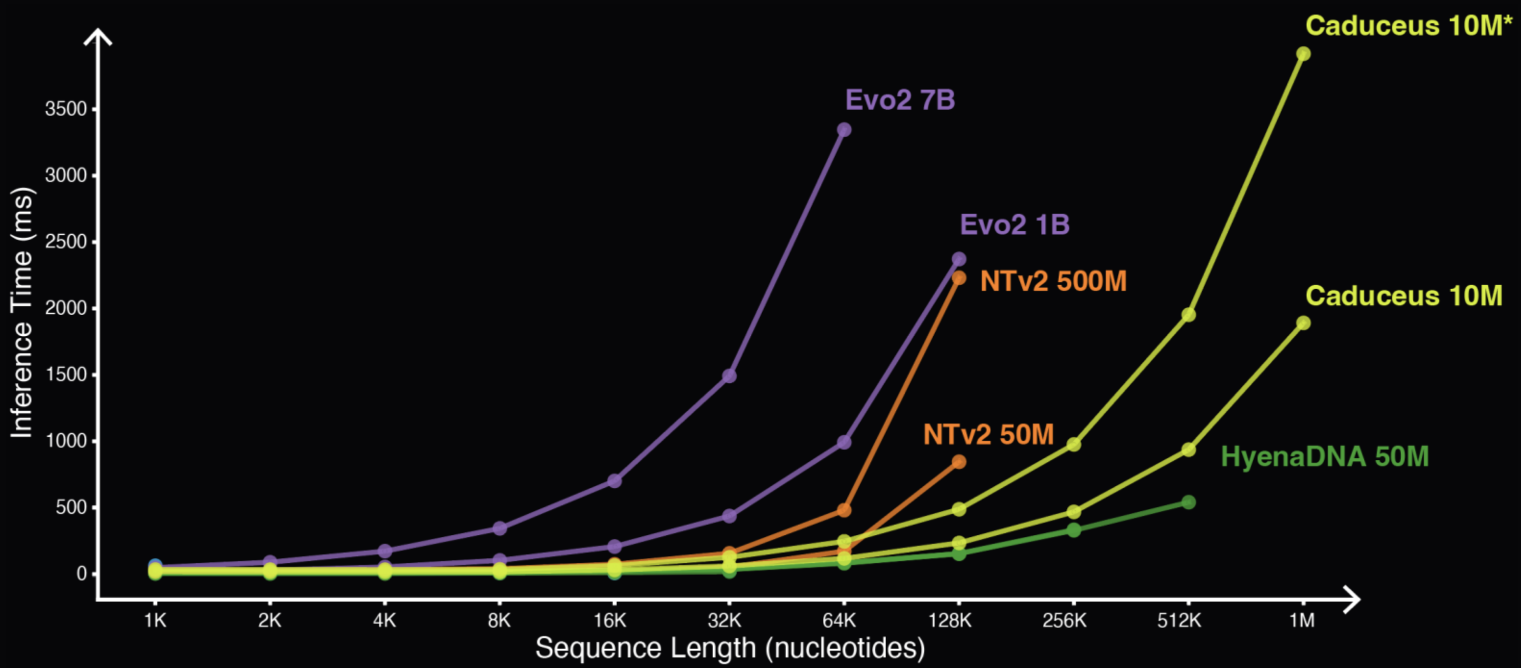


# Pre-training | Scaling Laws in Action



# Fine-tuning | NTv3, The Fastest Genomic Foundation Models

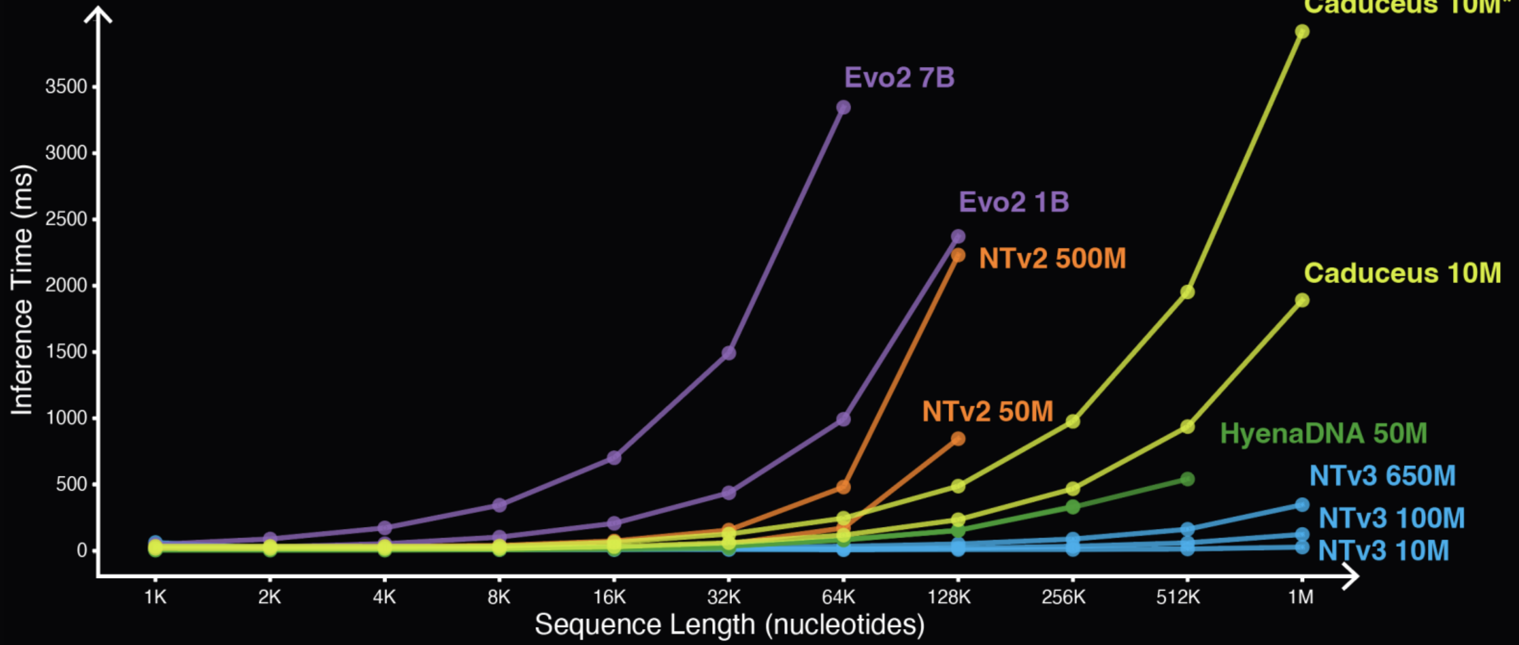
NTv3 scales up to 1 million nucleotides an order of magnitude more efficiently than competitive models



*RC equivariant

# Fine-tuning | NTv3, The Fastest Genomic Foundation Models

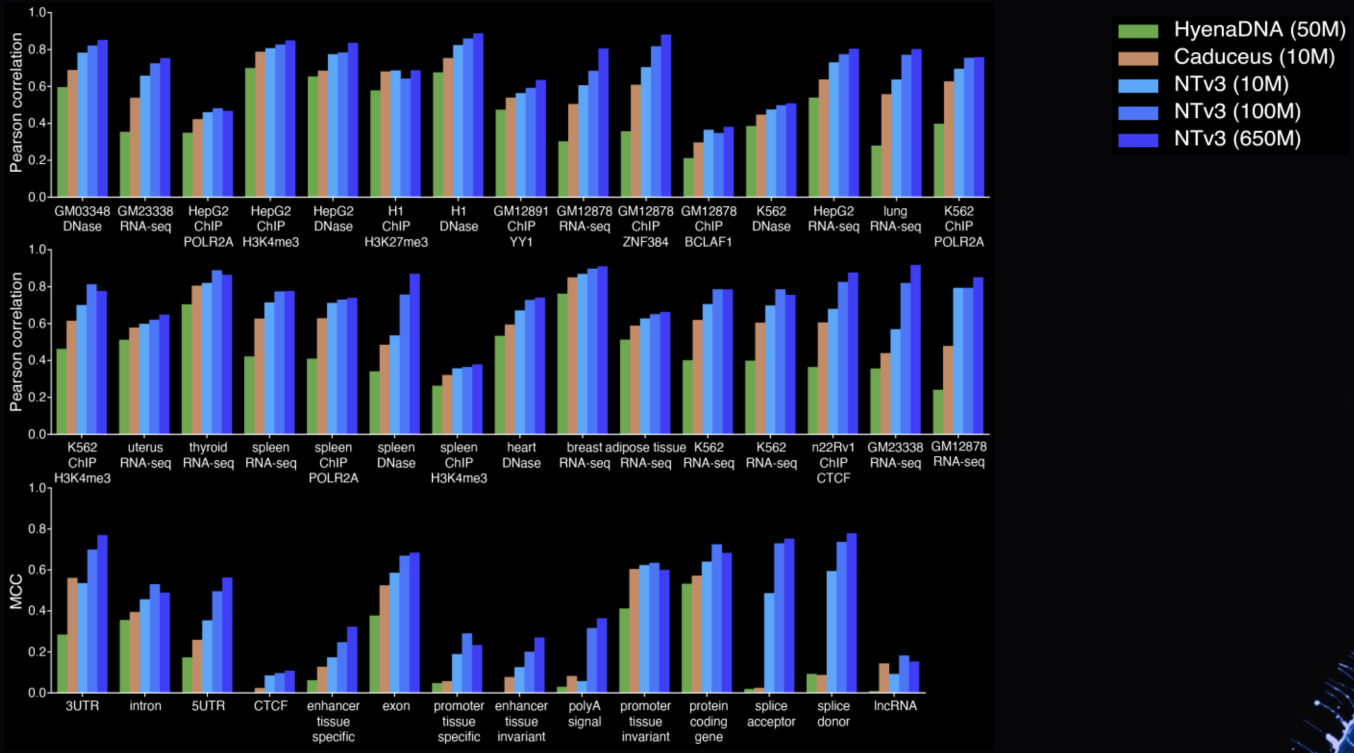
NTv3 scales up to 1 million nucleotides an order of magnitude more efficiently than competitive models



*RC equivariant

# Fine-tuning | NTv3 is amongst the best models for fine-tuning on downstream tasks

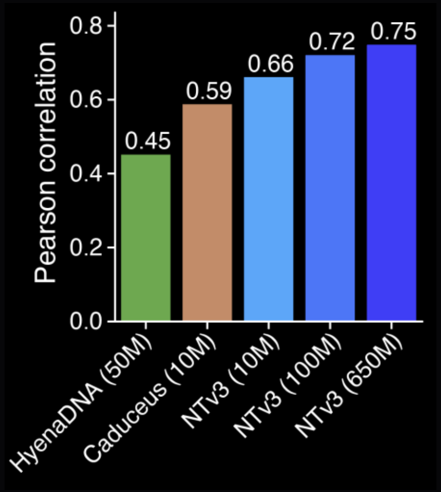
Evaluation of different foundation models on 44 long-range downstream tasks, including gene expression, DNA accessibility and genome annotation across various human tissues.



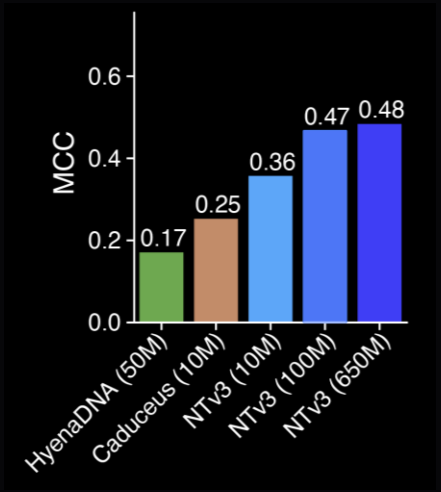
# Fine-tuning | NTv3 is amongst the best models for fine-tuning on downstream tasks

Evaluation of different foundation models on 44 long-range downstream tasks, including gene expression, DNA accessibility and genome annotation across various human tissues.

Average performance across quantitative tasks



Average performance across classification tasks

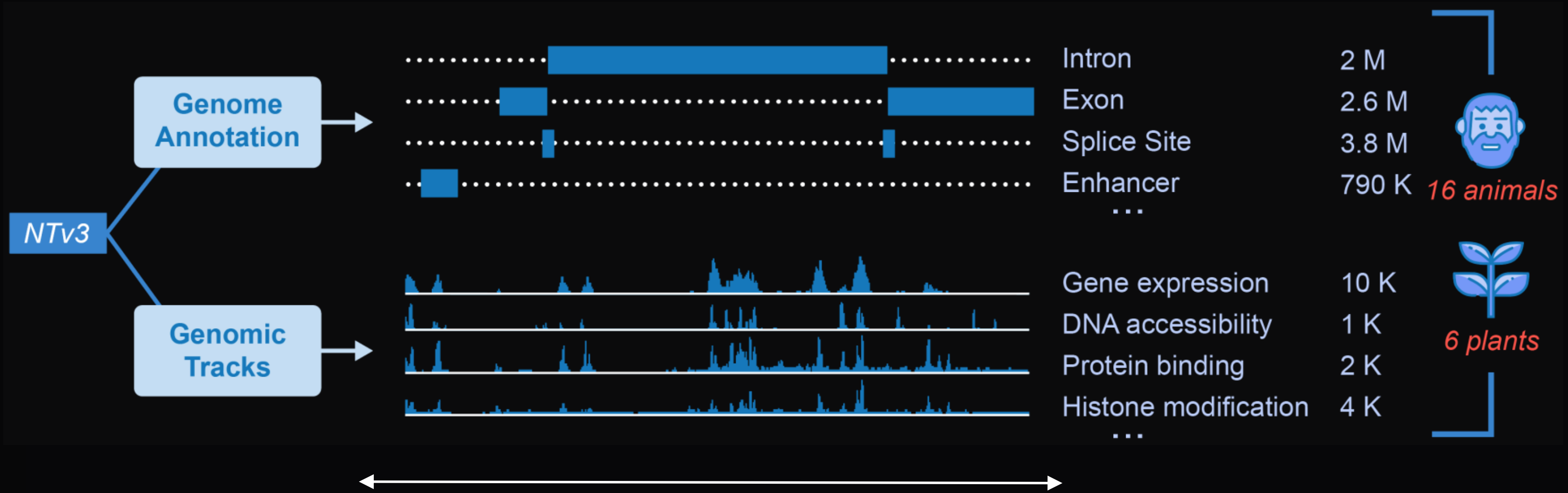


## NTv3

Best small foundation model (10M)

Top performance with larger model size

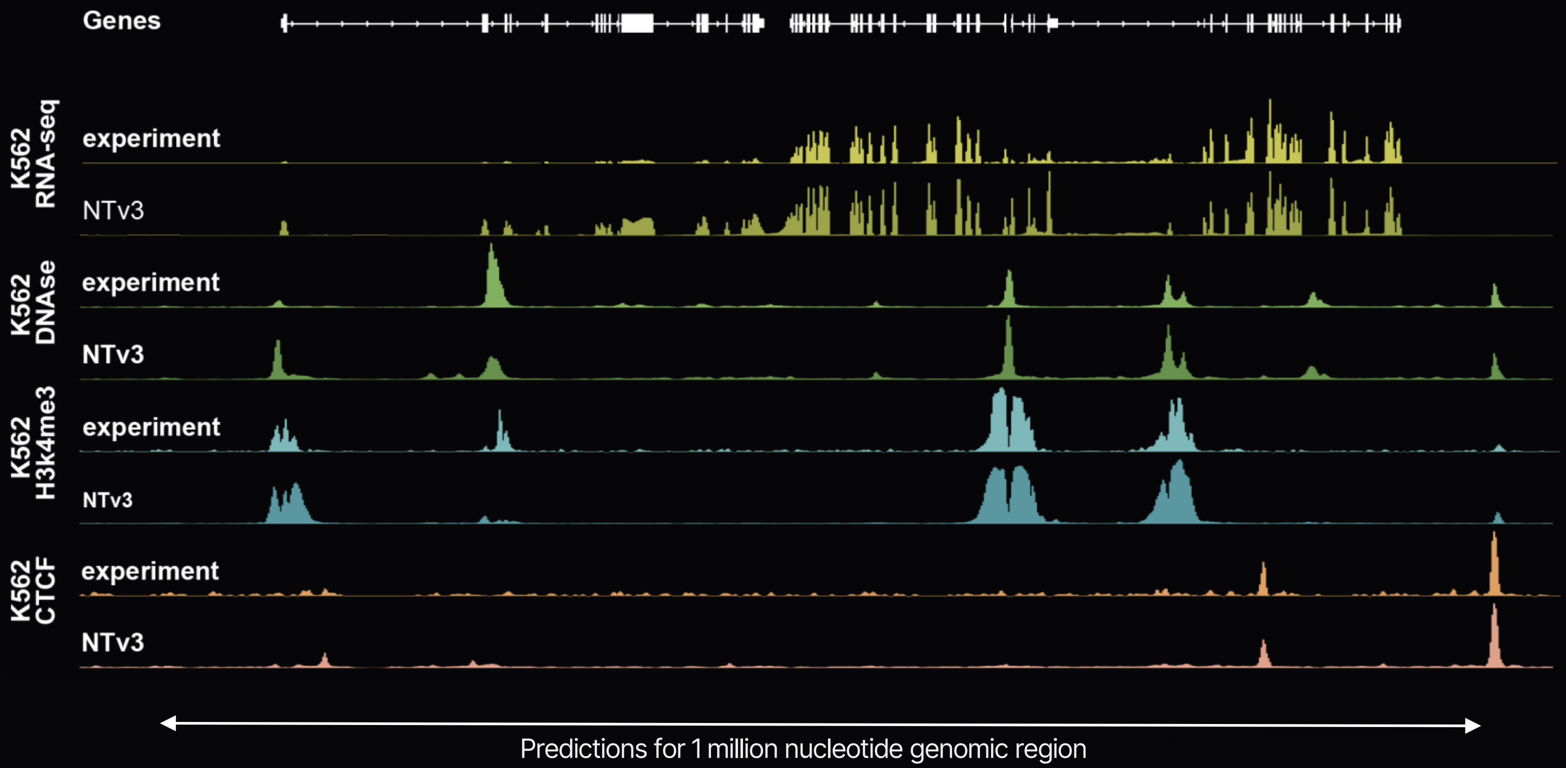
# Post-training | Learning from +17k genomic tracks and genome annotation



Predictions for 1mb input sequence
at single-nucleotide resolution

# Post-training | NTv3 accurately predicts genomic tracks at single-nucleotide resolution

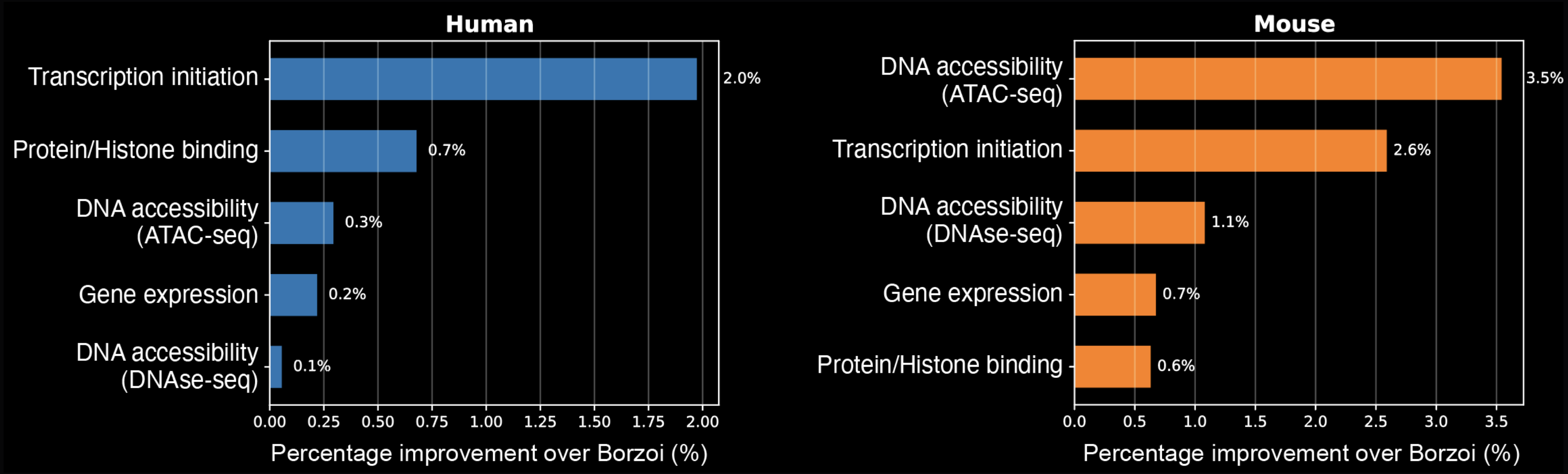
Example of NTv3 predictions for experiments in human K562 leukemia cells



Predictions for 1 million nucleotide genomic region

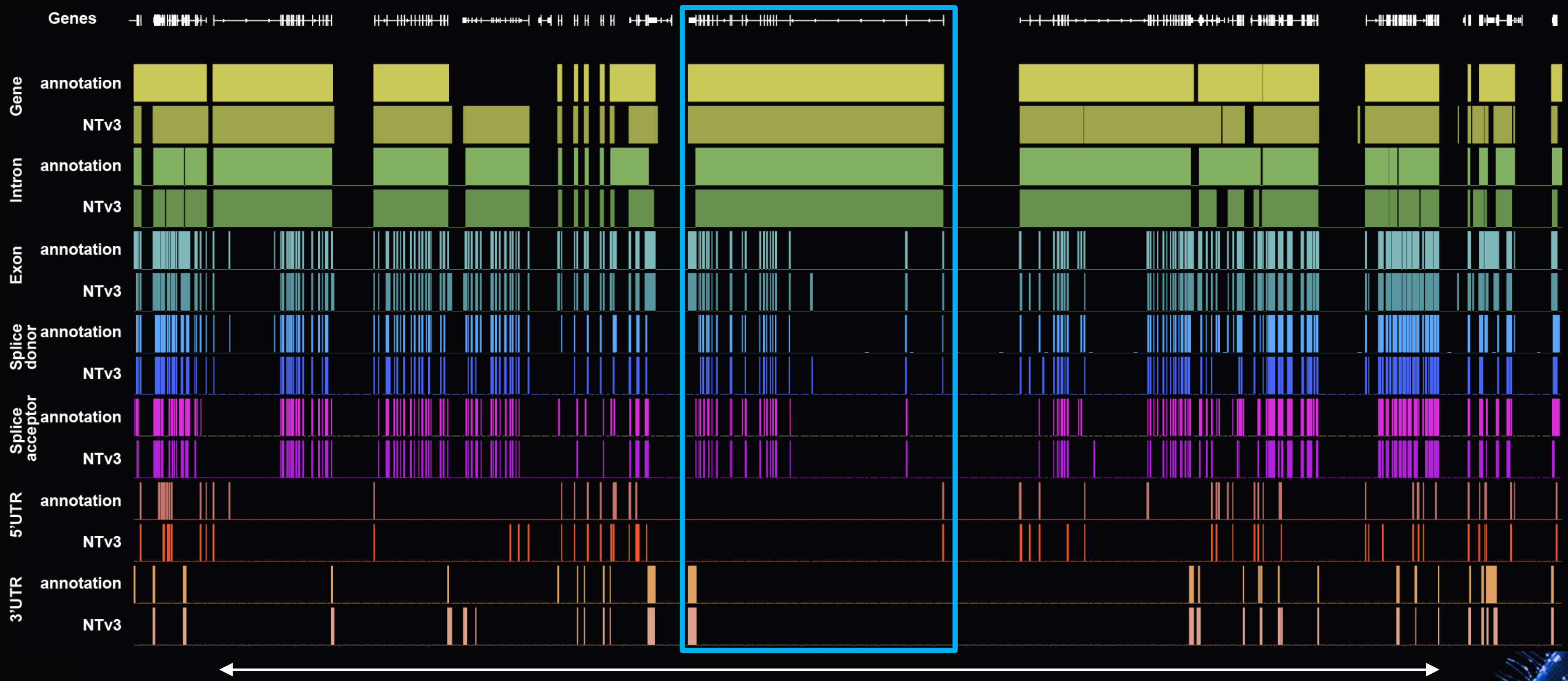
# Post-training | NTv3 accurately predicts genomic tracks at single-nucleotide resolution

NTv3 outperforms state-of-the-art model (Borzoi*) at experimental track prediction.



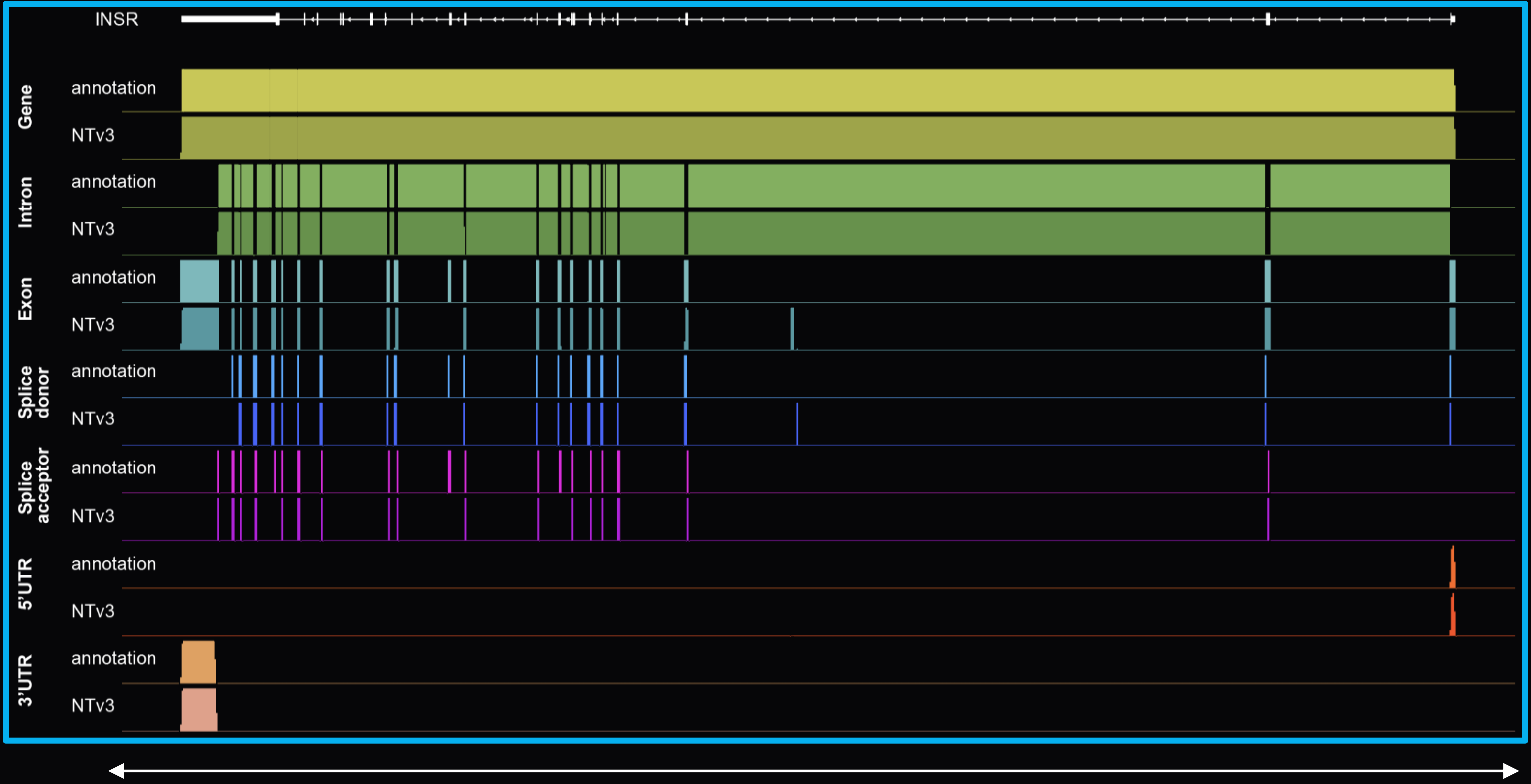
*Linder et al., Nature Genetics 2025

# Post-training | NTv3 accurately annotates genomes at single-nucleotide resolution



Predictions for 1 million nucleotide genomic region

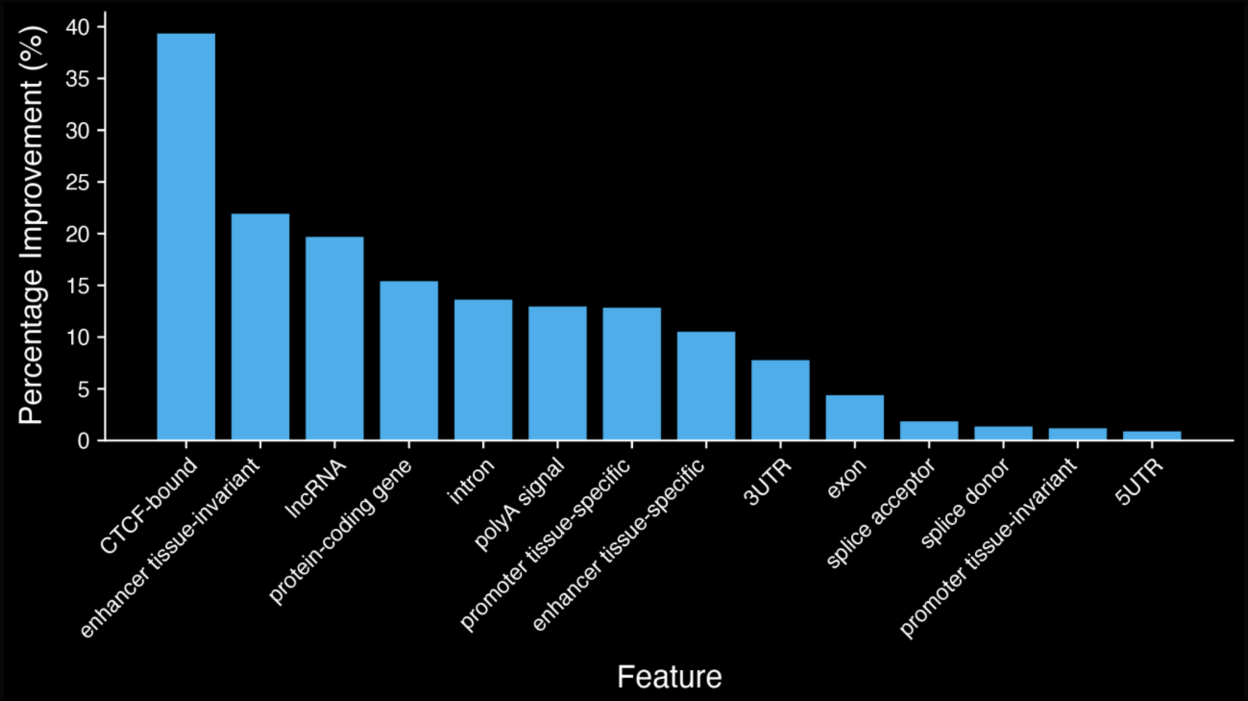
# Post-training | NTv3 accurately annotates genomes at single-nucleotide resolution



Predictions for 200 thousand nucleotide gene region

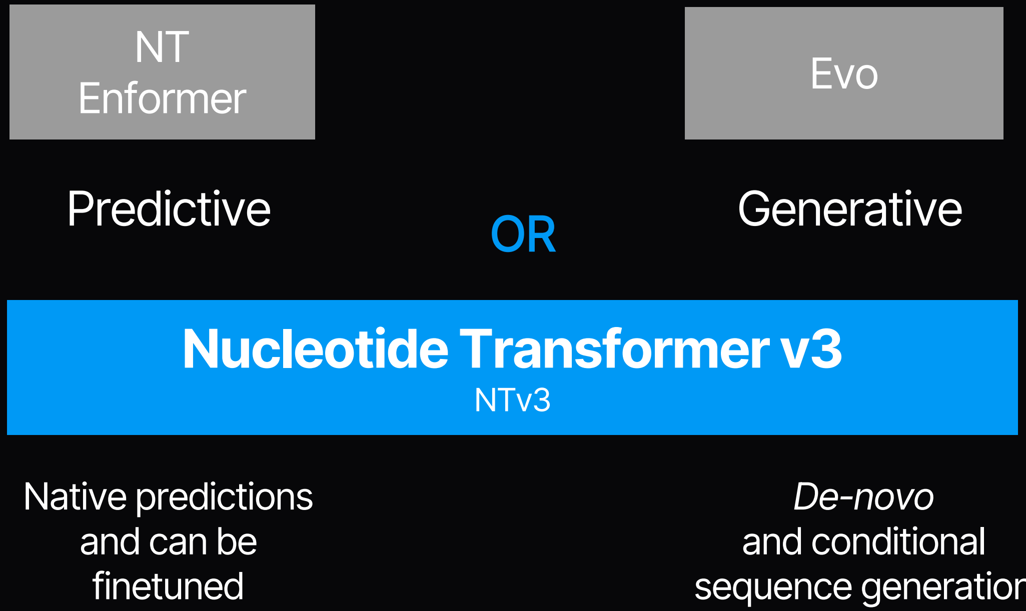
# Post-training | NTv3 accurately annotates genomes at single-nucleotide resolution

NTv3 outperforms state-of-the-art model (SegmentNT*) at gene finding, regulatory element detection and splicing.



*de Almeida et al., Nature Methods 2025

# Exploiting the data available with the aim of building a best-in-class model for genomics



NT
Enformer

Predictive

OR

Evo

Generative

**Nucleotide Transformer v3**
NTv3

Native predictions
and can be
finetuned

*De-novo*
and conditional
sequence generation

Thanks to the masked discrete diffusion framework, NTv3 both exhibits strong representation capabilities for downstream tasks and is generative!

# Generation | Designing regulatory enhancer sequences with NTv3

## Experiment

Design promoter-specific enhancers, at different levels of activity, in *Drosophila* cell line.

## Motivation

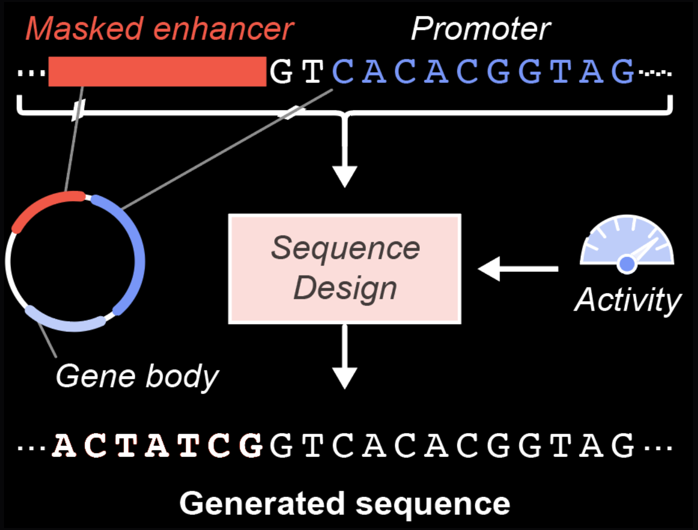
Enhancers are sequence elements that modulate the expression of genes and can be used for gene therapy.

## Approach

Fine-tune NTv3 to become a generative model using Masked Diffusion Language Models (MDLM)

## Validation

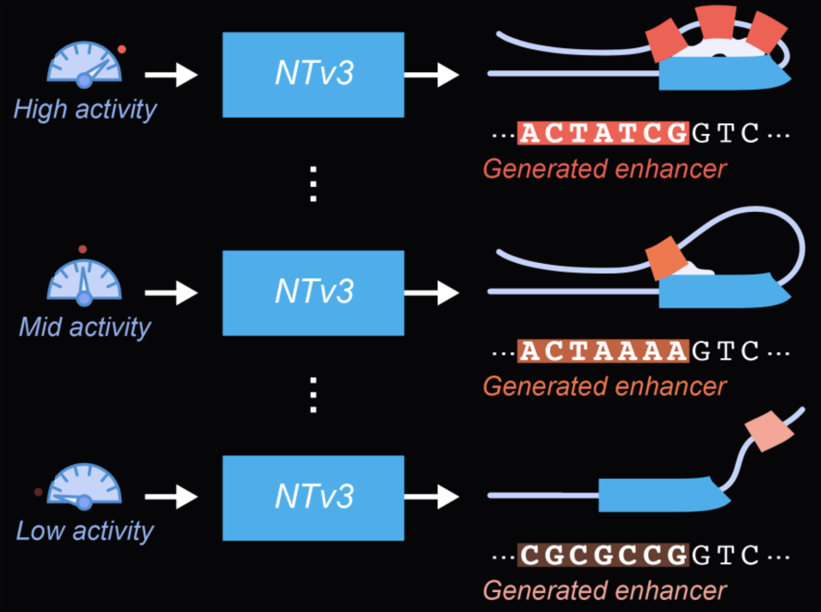
Experimental validation through *in vitro* MPRAs



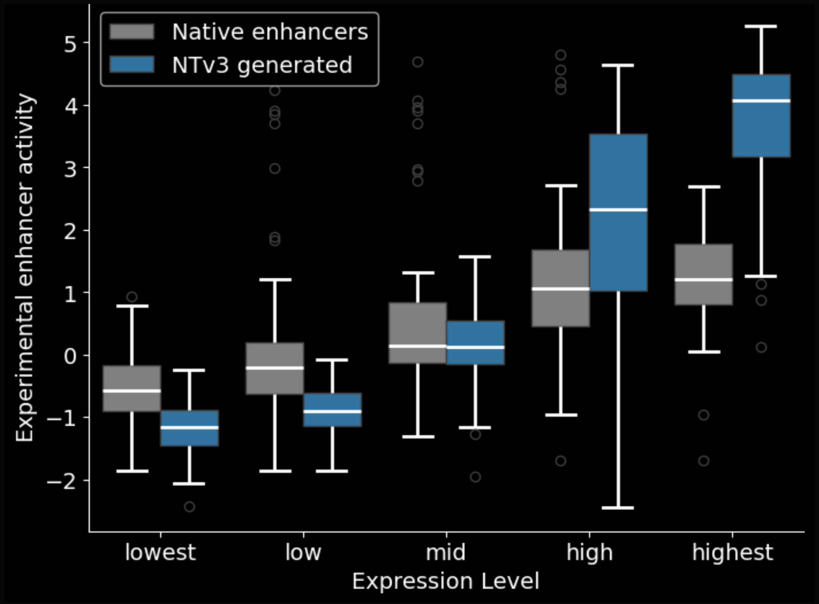
*In collaboration with Alex Stark*

IMP Research Institute of Molecular Pathology

# Generation | NTv3 designs have *in-vitro* state-of-the-art performance for activity-specific design

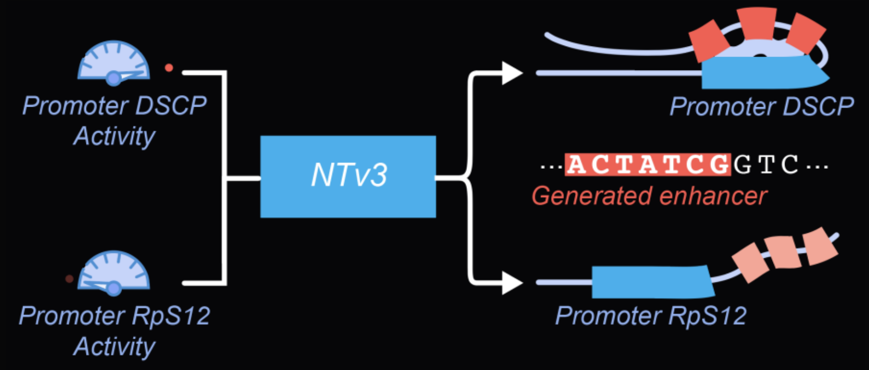


Experimental validation of enhancers
with different strengths (RpS12 promoter)

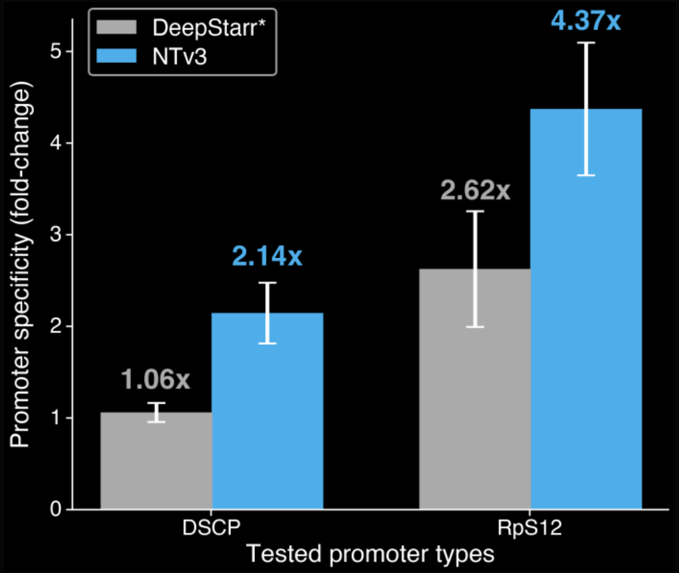


NTv3 successfully generated *de novo* enhancers matching the prompted activity levels.

# Generation | NTv3 designs have *in-vitro* state-of-the-art performance for promoter-specific design



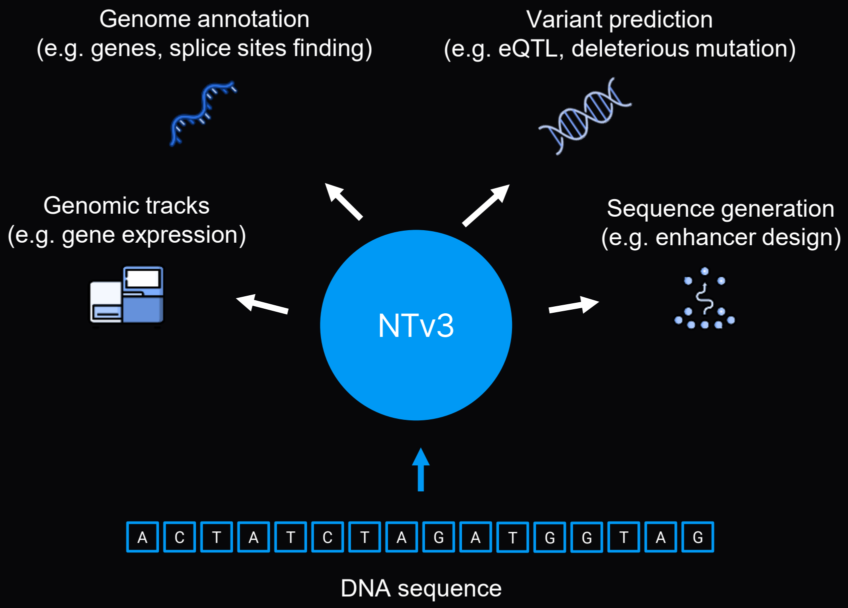
Experimental validation of enhancers with promoter-specific activities



NTv3 successfully generated *de novo* promoter-specific enhancers, achieving fold-change specificity significantly superior to previously validated state-of-the-art *in vitro* methods.

*de Almeida et al., Nature Genetics 2023

# NTv3: a new generation foundational model for genomics applications



Genome annotation
(e.g. genes, splice sites finding)

Variant prediction
(e.g. eQTL, deleterious mutation)

Genomic tracks
(e.g. gene expression)

Sequence generation
(e.g. enhancer design)

NTv3

A C T A T C T A G A T G G T A G

DNA sequence

AI Innovation

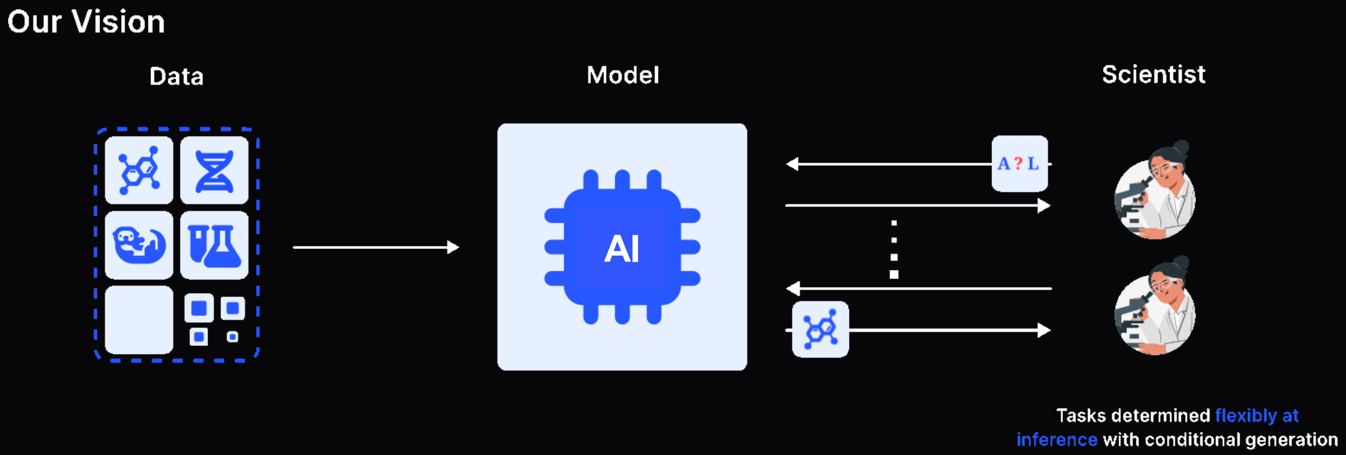
# Generative AI for protein and antibody engineering

Bora Guloglu
Senior Research Scientist

# Our vision: one model, many tasks

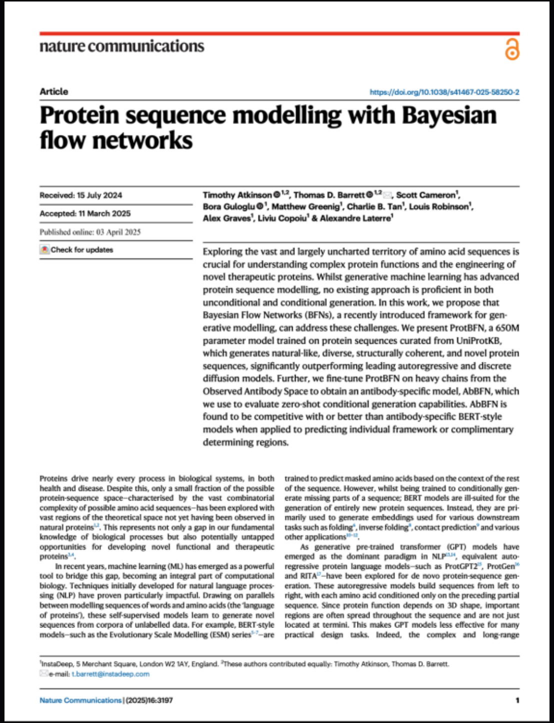
Our goal is to model as much of the data as possible

- Superior performance by learning a joint distribution across multiple data types and sources.
- Unparalleled flexibility in the hands of scientists with task-specific inference.



Tasks determined flexibly at inference with conditional generation

# Laying the Groundwork



nature communications

Article https://doi.org/10.1038/s41467-025-58250-2

## Protein sequence modelling with Bayesian flow networks

Received: 15 July 2024

Accepted: 11 March 2025

Published online: 03 April 2025

Check for updates

Timothy Atkinson[1,2], Thomas D. Barrett[1,2] ✉, Scott Cameron[1], Bora Guloglu[1], Matthew Greenig[1], Charlie B. Tan[1], Louis Robinson[1], Alex Graves[1], Liviu Copoiu[1] & Alexandre Laterre[1]

Exploring the vast and largely uncharted territory of amino acid sequences is crucial for understanding complex protein functions and the engineering of novel therapeutic proteins. Whilst generative machine learning has advanced protein sequence modelling, no existing approach is proficient in both unconditional and conditional generation. In this work, we propose that Bayesian Flow Networks (BFNs), a recently introduced framework for generative modelling, can address these challenges. We present ProtBFN, a 650M parameter model trained on protein sequences curated from UniProtKB, which generates natural-like, diverse, structurally coherent, and novel protein sequences, significantly outperforming leading autoregressive and discrete diffusion models. Further, we fine-tune ProtBFN on heavy chains from the Observed Antibody Space to obtain an antibody-specific model, AbBFN, which we use to evaluate zero-shot conditional generation capabilities. AbBFN is found to be competitive with or better than antibody-specific BERT-style models when applied to predicting individual framework or complimentary determining regions.

Proteins drive nearly every process in biological systems, in both health and disease. Despite this, only a small fraction of the possible protein-sequence space—characterised by the vast combinatorial complexity of possible amino acid sequences—has been explored with vast regions of the theoretical space not yet having been observed in natural proteins[1,2]. This represents not only a gap in our fundamental knowledge of biological processes but also potentially untapped opportunities for developing novel functional and therapeutic proteins[3,4].

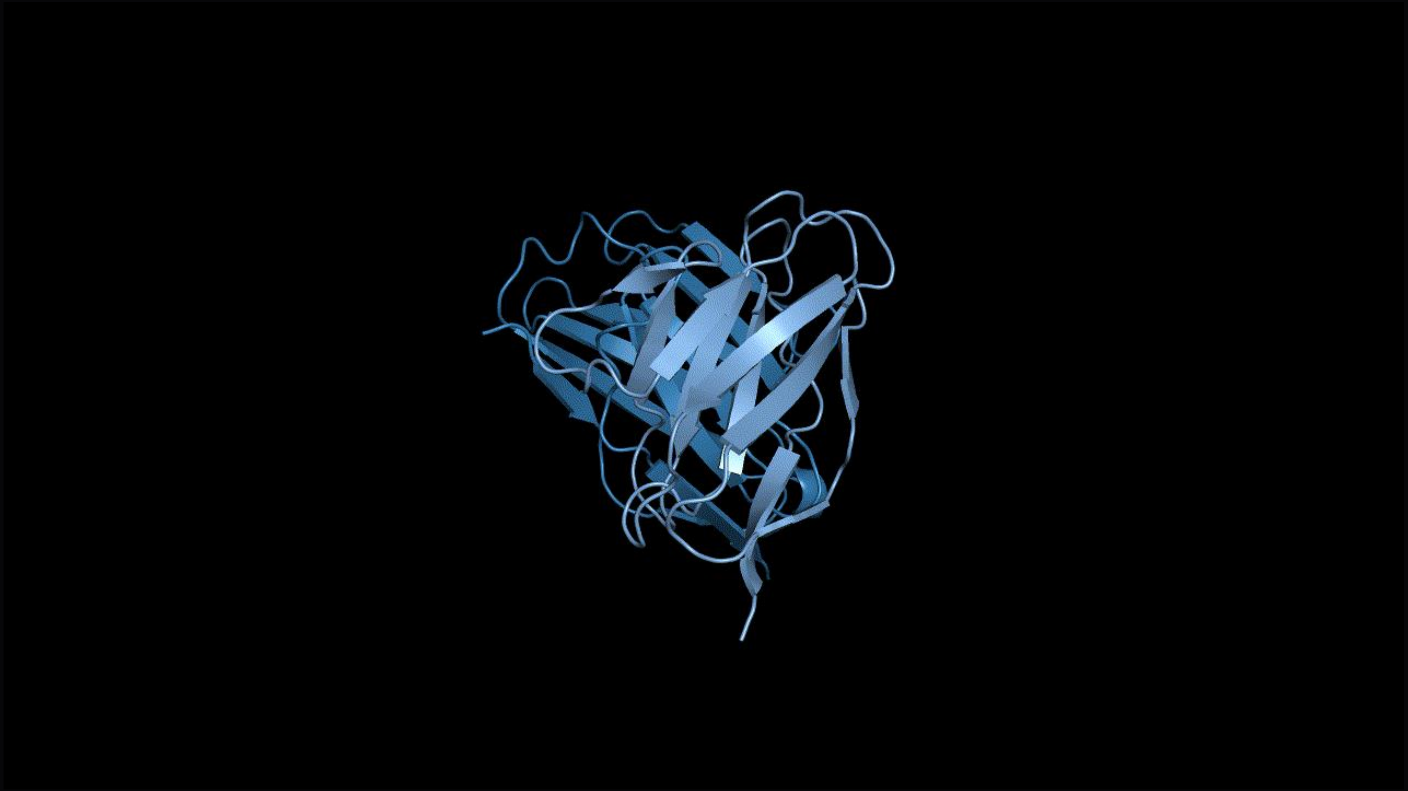
In recent years, machine learning (ML) has emerged as a powerful tool to bridge this gap, becoming an integral part of computational biology. Techniques initially developed for natural language processing (NLP) have proven particularly impactful. Drawing on parallels between modelling sequences of words and amino acids (the 'language of proteins'), these self-supervised models learn to generate novel sequences from corpora of unlabelled data. For example, BERT-style models—such as the Evolutionary Scale Modelling (ESM) series[5–7]—are trained to predict masked amino acids based on the context of the rest of the sequence. However, whilst being trained to conditionally generate missing parts of a sequence; BERT models are ill-suited for the generation of entirely new protein sequences. Instead, they are primarily used to generate embeddings used for various downstream tasks such as folding[6], inverse folding[8], contact prediction[9] and various other applications[10–12].

As generative pre-trained transformer (GPT) models have emerged as the dominant paradigm in NLP[13,14], equivalent autoregressive protein language models—such as ProtGPT2[15], ProtGen[16] and RITA[17]—have been explored for de novo protein-sequence generation. These autoregressive models build sequences from left to right, with each amino acid conditioned only on the preceding partial sequence. Since protein function depends on 3D shape, important regions are often spread throughout the sequence and are not just located at termini. This makes GPT models less effective for many practical design tasks. Indeed, the complex and long-range

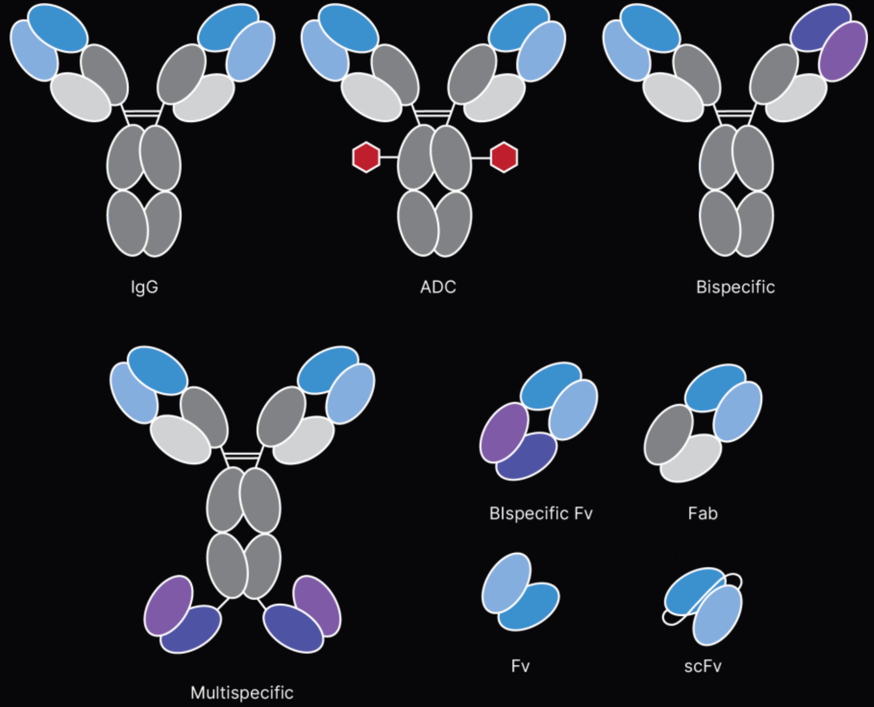
[1] InstaDeep, 5 Merchant Square, London W2 1AY, England. [2] These authors contributed equally: Timothy Atkinson, Thomas D. Barrett. ✉ e-mail: t.barrett@instadeep.com

1. Atkinson, T., Barrett, T.D., Cameron, S. *et al.* Protein sequence modelling with Bayesian flow networks. *Nat Commun* **16**, 3197 (2025)

# AbBFN2



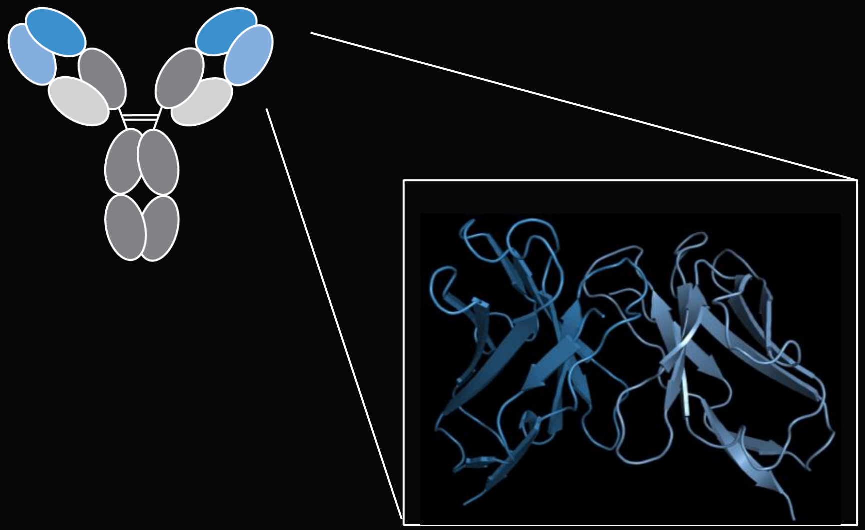
# Why design $F_v$s?



Therapeutic antibodies have diversified in formats across the years.

The $F_v$ region is the common key recognition component in all modalities.

# AbBFN2



Antibodies have unusual properties, necessitating fine-grained control over their genetics, sequence, and overall biophysics.

It is estimated that $>10^{16}$ antibody sequences are possible: needle-in-a-haystack problem with multiple design objectives.

CDR-H1 CDR-H2 CDR-H3

*VH*: EVQLLESGGGLVQPGGSLRLSCAAS**GFTFSSYA**MSWVRQAPGKGLEWVSA**ISWNSGSI**YADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYC**ARGWSQVDTAMDLDY**GQGTLVTVSS

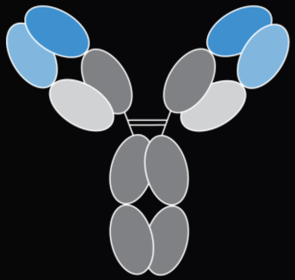
*V gene* *D gene* *J gene*

CDR-L1 CDR-L2 CDR-L3

*VL*: DIQMTQSPSSVSASVGDRVTITCRASQSVSSNLAWYQQKPGKAPKLLIYGASSLQSGVPSRFSGSGSGTDFTLTISSLQPEDFATYYCQQYNNWLTFGQGTRLEIK

*V gene* *J gene*

# AbBFN2: included data modalities



## Biophysical Attributes

- Hydrophobicity
- Positive Patches
- Negative Patches
- Charge Imbalance
- Hydrophobicity Flag
- Positive Patches Flag
- Negative Patches Flag
- Charge Imbalance Flag

## Amino Acid Sequence

- AGL • FWR-H1
- AGL • CDR-H1
- AGL • FWR-H2
- AGL • CDR-H2
- AGL • FWR-H3
- AGL • CDR-H3
- AGL • FWR-H4
- AGL • FWR-L1
- AGL • CDR-L1
- AGL • FWR-L2
- AGL • CDR-L2
- AGL • FWR-L3
- AGL • CDR-L3
- AGL • FWR-L4

## Genetic Attributes

- HV gene
- HD gene
- HU gene
- HV seq identity
- HD seq identity
- HU seq identity
- HV seq identity
- HD seq identity
- HJ seq identity
- LV gene
- LJ gene
- LV seq identity
- LJ seq identity
- LV seq identity
- LJ seq identity
- LC locus
- Species

## Length Attributes

- CDR-H1 length
- CDR-H2 length
- CDR-H3 length
- CDR-L1 length
- CDR-L2 length
- CDR-L3 length

## Novel capabilities:

- Per-residue energies
- Vernier zone energies
- Interface energies
- Germline families
- Per-residue genetics

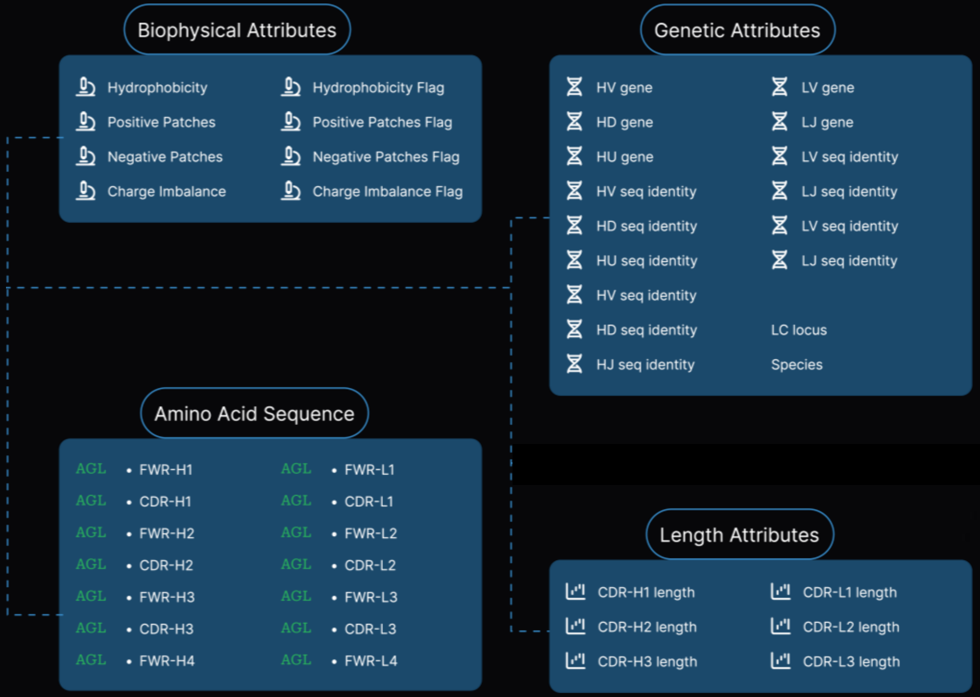
## Future Plans

- Antibody Structure
- Antigen Structure
- Quantum-level energetics
- Arbitrary assay data
- Binding interactions

# Sequence annotation

AbBFN2 achieves SOTA results on 23/23 sequence labelling tasks[1], demonstrating robust learning of the genetic and biophysical attributes of antibody sequences.

- Sequence labelling is a prerequisite for steered generation and design.
- AbBFN2 is a one-stop labelling tool,
- Simplifies traditional computational pipelines and improves accuracy.



Biophysical Attributes

- Hydrophobicity
- Hydrophobicity Flag
- Positive Patches
- Positive Patches Flag
- Negative Patches
- Negative Patches Flag
- Charge Imbalance
- Charge Imbalance Flag

Genetic Attributes

- HV gene
- HD gene
- HU gene
- HV seq identity
- HD seq identity
- HU seq identity
- HV seq identity
- HD seq identity
- HJ seq identity
- LV gene
- LJ gene
- LV seq identity
- LJ seq identity
- LV seq identity
- LJ seq identity
- LC locus
- Species

Amino Acid Sequence

| | | | |
|---|---|---|---|
| AGL | • FWR-H1 | AGL | • FWR-L1 |
| AGL | • CDR-H1 | AGL | • CDR-L1 |
| AGL | • FWR-H2 | AGL | • FWR-L2 |
| AGL | • CDR-H2 | AGL | • CDR-L2 |
| AGL | • FWR-H3 | AGL | • FWR-L3 |
| AGL | • CDR-H3 | AGL | • CDR-L3 |
| AGL | • FWR-H4 | AGL | • FWR-L4 |

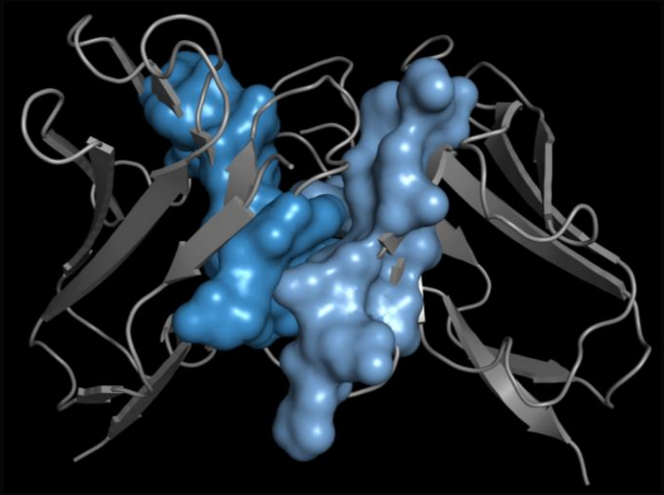
Length Attributes

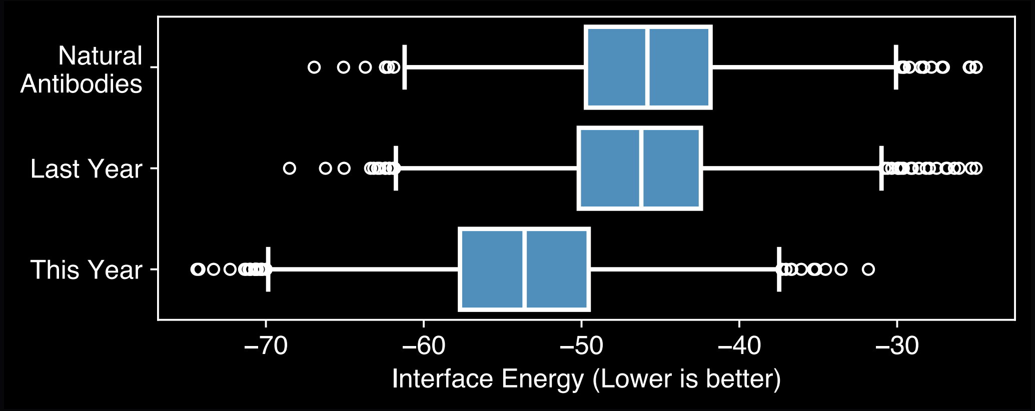
- CDR-H1 length
- CDR-H2 length
- CDR-H3 length
- CDR-L1 length
- CDR-L2 length
- CDR-L3 length

[1] Experiment conducted on 10,000 unseen antibodies which were labelled by competitor models and traditional tools. Categorical data assessed via balanced F1 scores, continuous data assessed via Pearson's R and root mean squared error.

# Stabilisation of existing antibodies

Using an unstable starting candidate, AbBFN2 is able to refine the interfaces and the total antibody to increase stability[1], which allows more stable pairing, better storage, and higher expression levels.

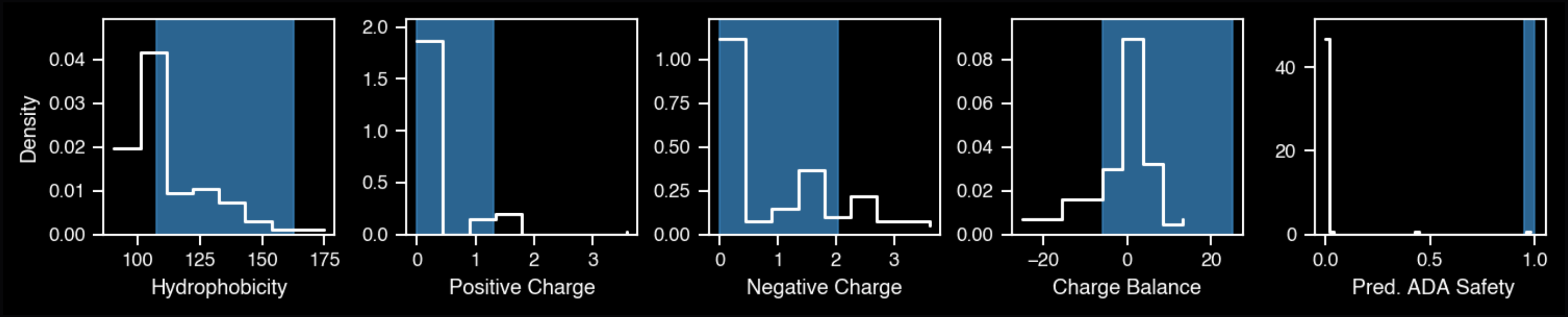




[1] Interface energies are calculated using the Rosetta Protein Modelling Suite. 5,000 samples are generated in each case and compared to a background distribution 5,000 randomly picked unseen antibodies.

# Multi-objective design using AbBFN2

- AbBFN2 optimises sequences with multiple conditions[1], using inference time compute scaling to generate diverse candidates for early discovery and optimisation.



### Inference-time compute scaling

EVQLLESGGGLVQPGGSLRLSCAAS...

QVQLLESGGSLVQPGGSLRLSCAAS...

QVQLLESGGSLVQPGGSIRLSCARS...

**>80%:** Success rate (overall)

**>90%:** Success rate (tractable candidates)

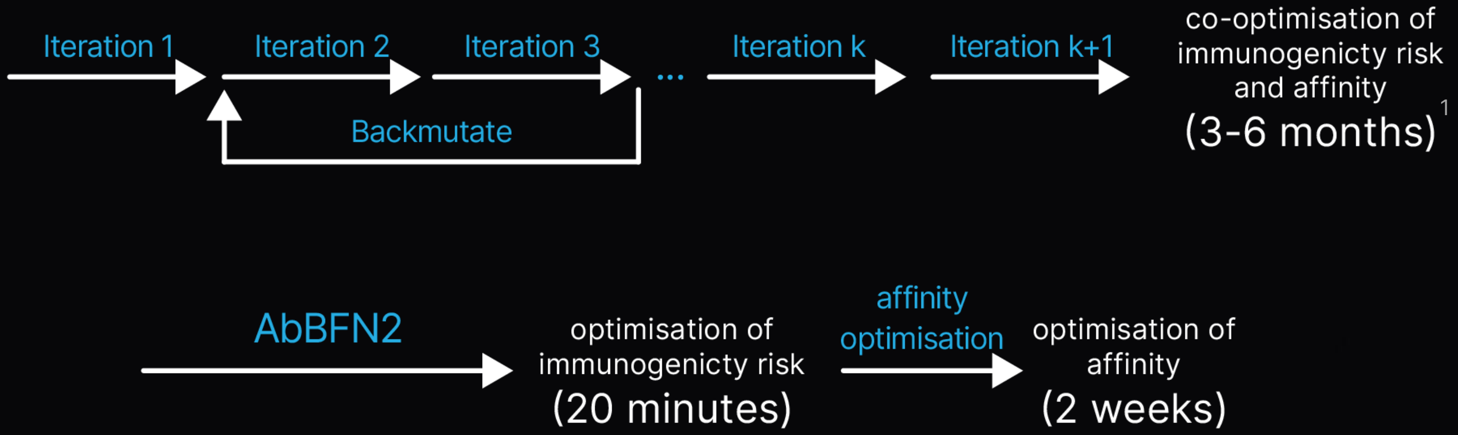
**46.6:** Number of mutations (1 objective)

**56.9:** Number of mutations (5 objectives)

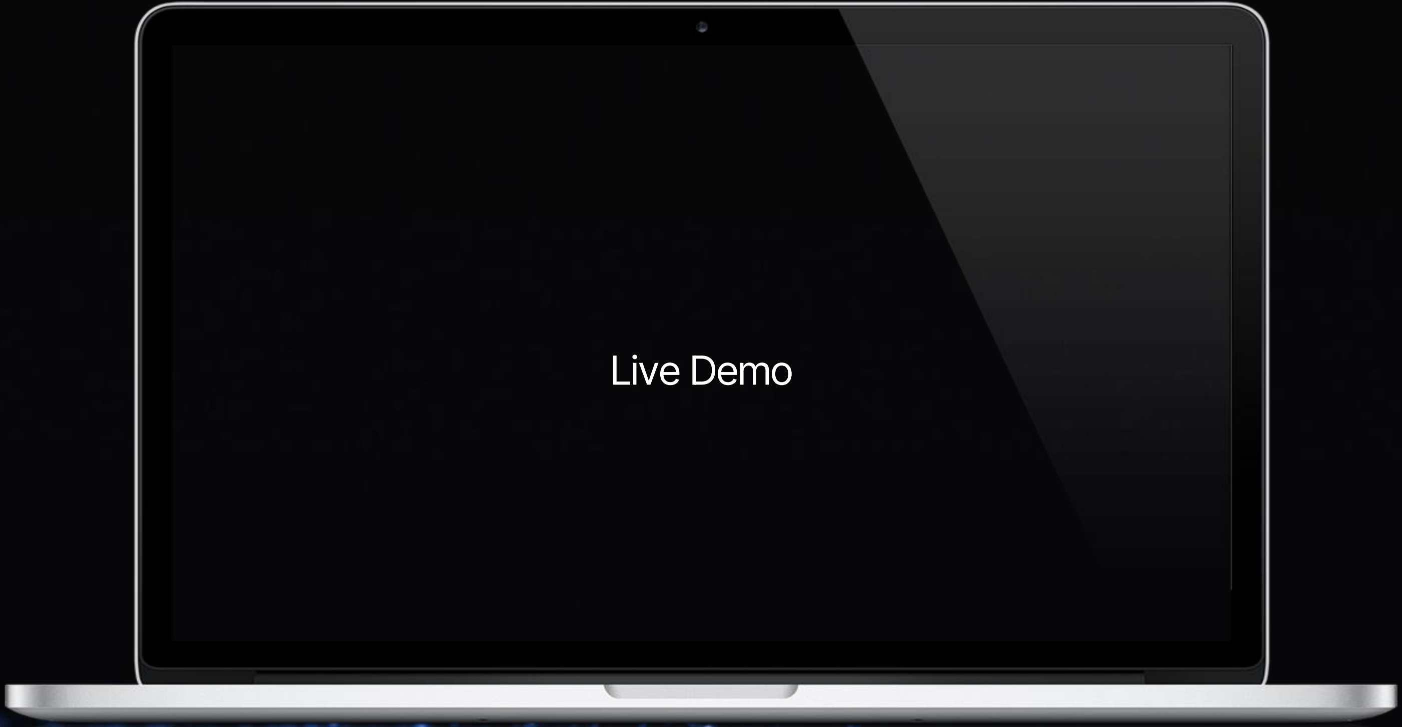
[1] 91 high-risk unseen sequences with multiple sequence liabilities were optimised. For each sequence, 4 candidates were generated with up to 15 recycling iterations. Results are reported for the best variant for each candidate.

# Experimental validation

Traditional humanization is a trial-and-error bottleneck: it often relies on arbitrary back-mutations, is time- and cost-intensive and risks disrupting binding through extensive changes.



[1] Based on Marks et al, *Bioinformatics*, 2021; Tennenhouse et al, *Nat. Biomed. Eng*, 2024



Live Demo

# Experimental validation

**AbBFN2 enables efficient *in silico* humanization, preserving antigen binding while eliminating the need for lengthy and expensive wet lab experiments.**

## Objective

Humanise antibodies to reduce side-effects, starting from the precursor sequence[1] to generate designs

## Results

Our sequences achieved similar expression with an average fewer edits compared to their manually designed clinical-stage counterparts.

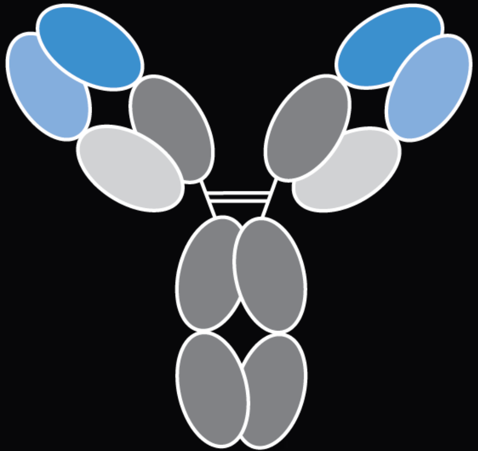
| Target | # mutations | | Binding (Kd, nM)[2] | |
|---|---|---|---|---|
| | Exp[3]. | Ours | Exp[3]. | Ours |
| IL-6Ra | **42** | **37** | 10.9 | **13.6** |
| IL-5 | 41 | **34** | 0.304 | **0.577** |
| Her2 | 55 | 63 | 14.1 | **29.3** |
| IgE | 60 | **42** | 2.82 | **7.92** |

1. For each precursor antibody, designs were generated as described previously. Designs were converted into scFv format and expressed in cell free *E. coli* expression systems. Expression levels are compared to the expression levels of the experimentally humanised reference sequence and binding was assessed using bio-layer interferometry. Experimentally humanised control sequences are those reported in *Marks et al, Bioinformatics, 2021*
2. Kd values might vary between reported and literature values due to experimental setup and selected scaffolds.
3. Exp. (Experimental) refers to a single molecule (per target) that is present in *Marks et al, Bioinformatics, 2021*

# AbBFN2: Cutting-edge in silico antibody design

- Training on both sequence and associated metadata of interest produces a rich syntax for "prompt/task engineering".

- The "condition anywhere, generate anywhere" paradigm of AbBFN2 admits a wide variety of tasks that can be decided at inference time.



# Data acquisition & refinement

Nicolas Lopez Carranza
Head of BioNTechAI
InstaDeep

Youssef Ben Dhieb
Senior ML Engineer
InstaDeep

# BioNTech AI strategy is **driven by data**

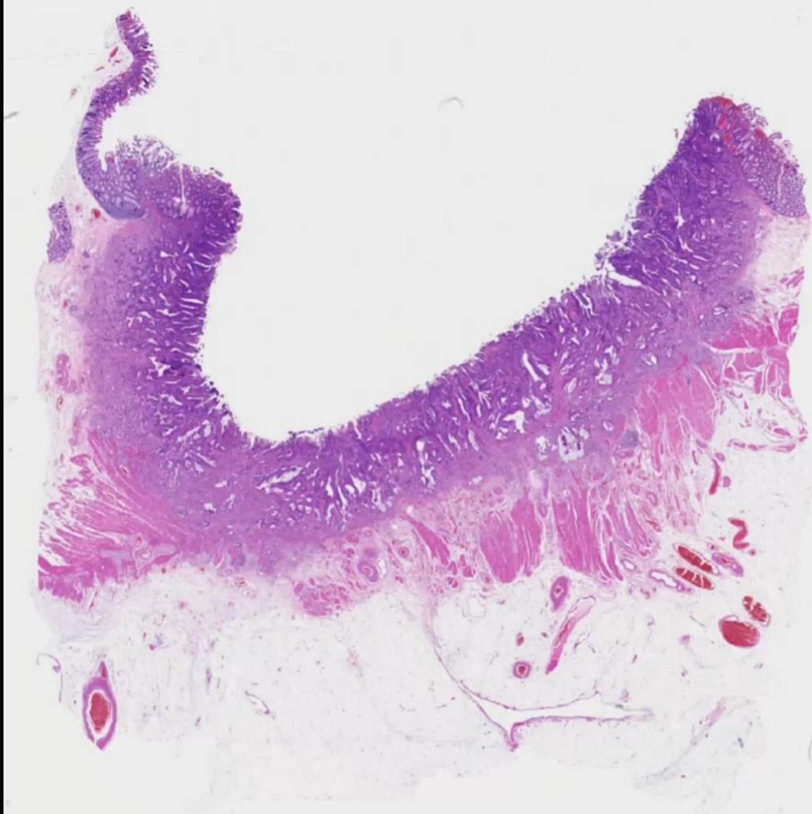
Potential for continued improvement as more data are generated and analysed

# We aim to learn as much as possible from the **tumour**

## Sequence Space

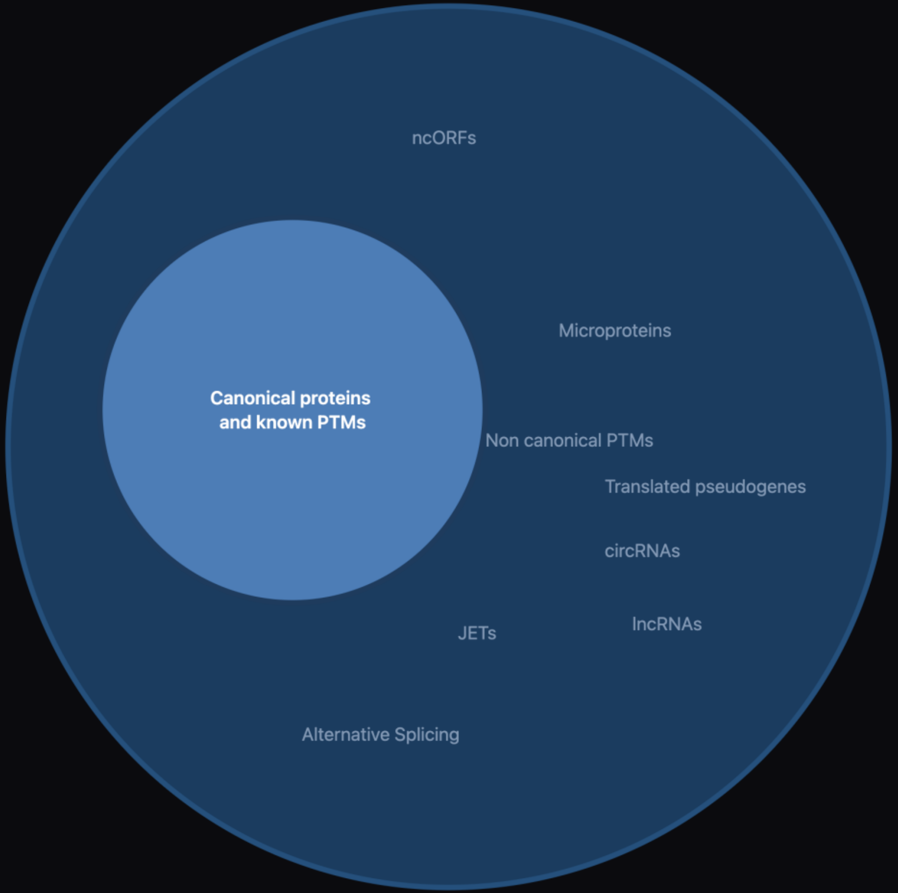
InstaNovo

## Image Space

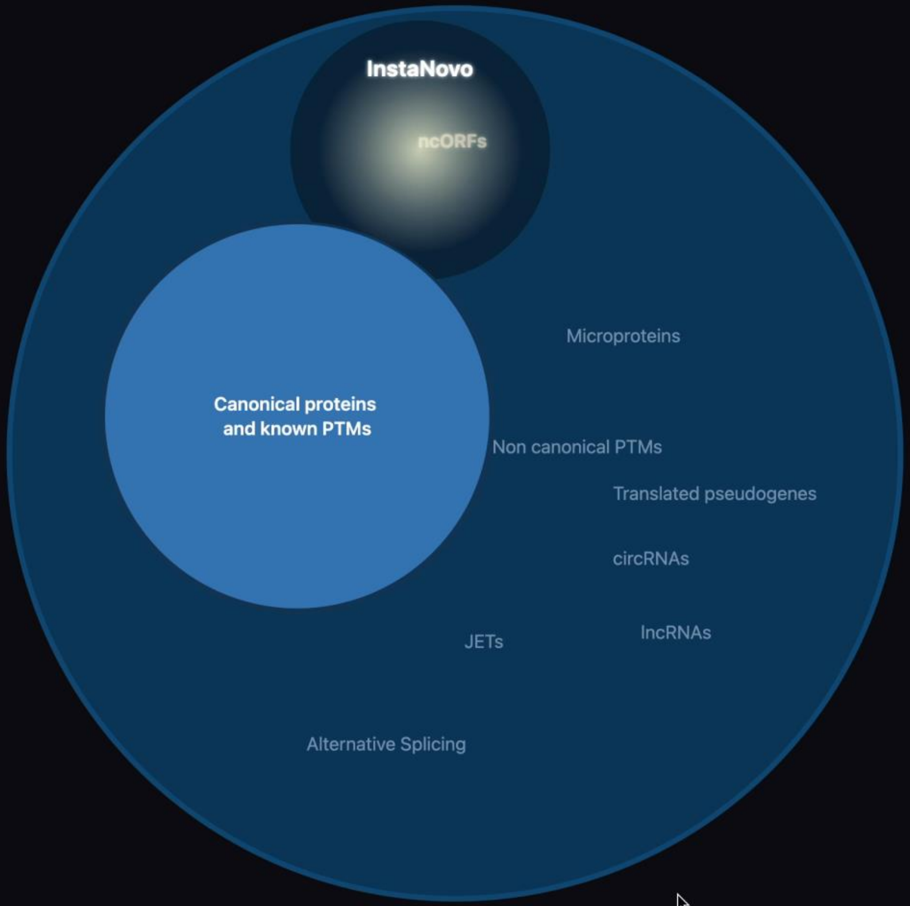


Internal

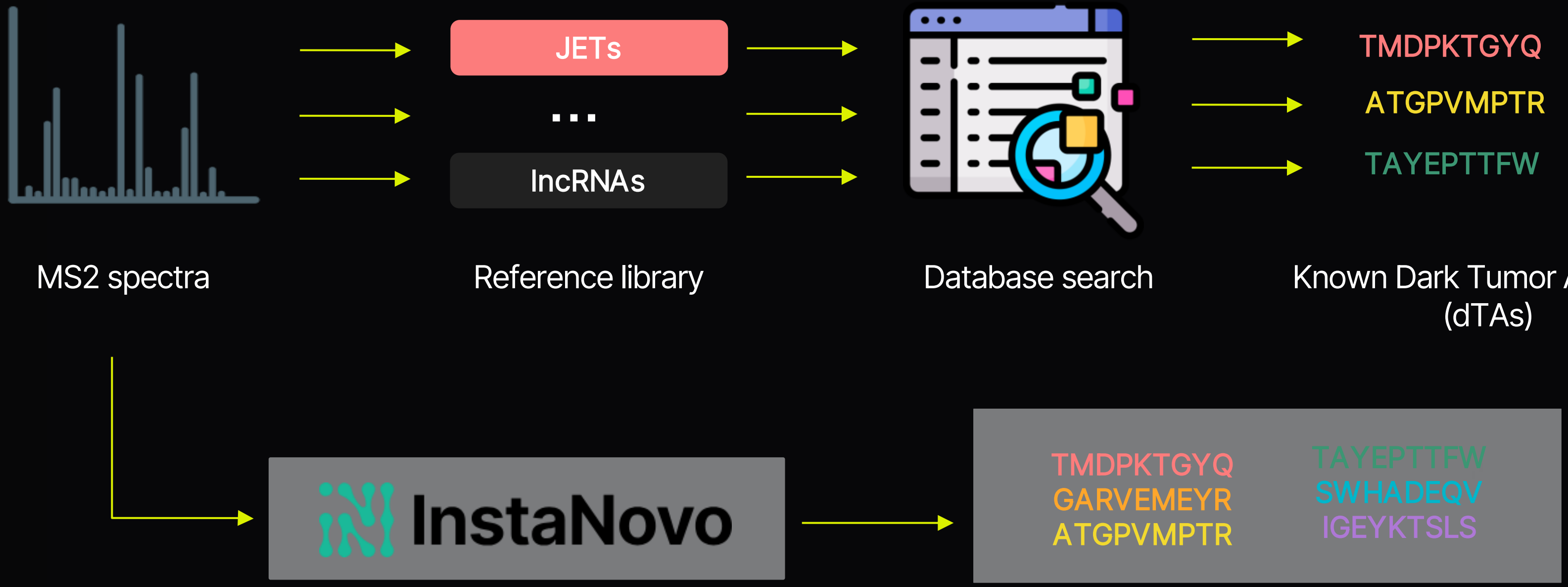
# The **Dark Proteome** encompasses uncharacterized proteins and hidden translation products beyond canonical proteins and known PTMs



# InstaNovo technology enables de novo peptide sequencing to explore the 'dark proteome' and uncover unknown proteins in cancer.



# InstaNovo’s library-free approach allows discovery of unanticipated dark proteome antigens



JETs

...

lncRNAs

TMDPKTGYQ

ATGPVMPTR

TAYEPTTFW

MS2 spectra

Reference library

Database search

Known Dark Tumor Antigens (dTAs)

InstaNovo

TMDPKTGYQ
GARVEMEYR
ATGPVMPTR

TAYEPTTFW
SWHADEQV
IGEYKTSLS

**All Possible Peptides**

*Incl. any unanticipated targets*

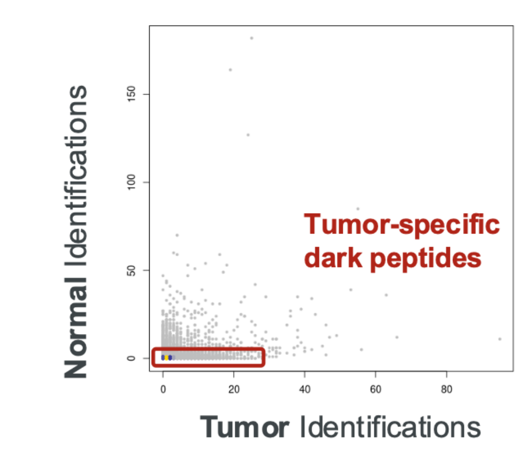
# InstaNovo SOTA de novo peptide sequencing

**InstaNovo** (auto-regressive) and **InstaNovo+** (diffusion) combine to outperform SOTA methods.

Has already shown potential **in detecting tumour specific epitopes** from undocumented ORFs and aberrant splicing.

Published in **Nature Machine Intelligence**

**Covered by Science Magazine**



| Peptide | Tumor | Normal |
|---|---|---|
| Peptide 1 | 32 | 1 |
| Peptide 2 | 28 | 1 |
| Peptide 3 | 26 | 1 |
| Peptide 4 | 31 | 0 |
| Peptide 5 | 28 | 1 |
| Peptide 6 | 30 | 0 |
| Peptide 7 | 26 | 1 |
| Peptide 8 | 27 | 1 |
| Peptide 9 | 27 | 1 |
| Peptide 10 | 27 | 1 |

nature machine intelligence

Article https://doi.org/10.1038/s42256-025-01019-5

**InstaNovo enables diffusion-powered de novo peptide sequencing in large-scale proteomics experiments**

Eloff, K., Kalogeropoulos, K., Mabona, A. et al. *InstaNovo enables diffusion-powered de novo peptide sequencing in large-scale proteomics experiments*. Nature Machine Intelligence 7, 565–579 (2025). https://doi.org/10.1038/s42256-025-01019-5

# Introducing InstaNovo V2

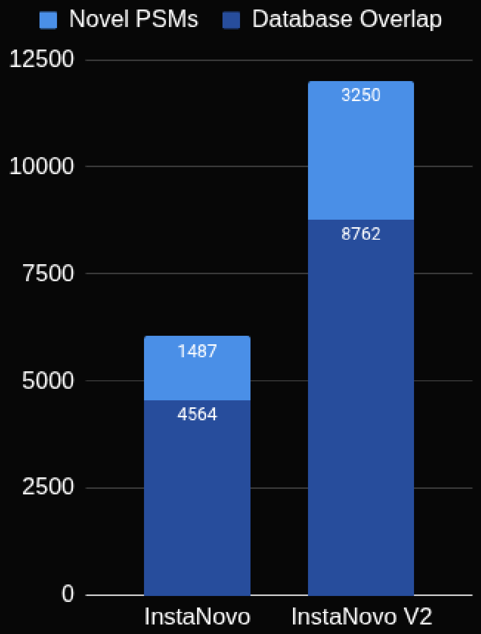
The next generation of InstaNovo models

**Larger Dataset**
63 million labelled spectra

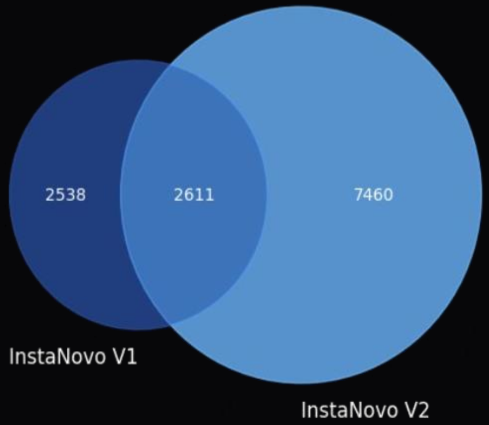
**Faster Prediction**
Up to 50x faster inference

**Higher Accuracy**
10–15% increase in peptide recovery

**More Identifications**
Up to 2× the number of identifications



Peptide identifications in InstaNovo V1 and V2

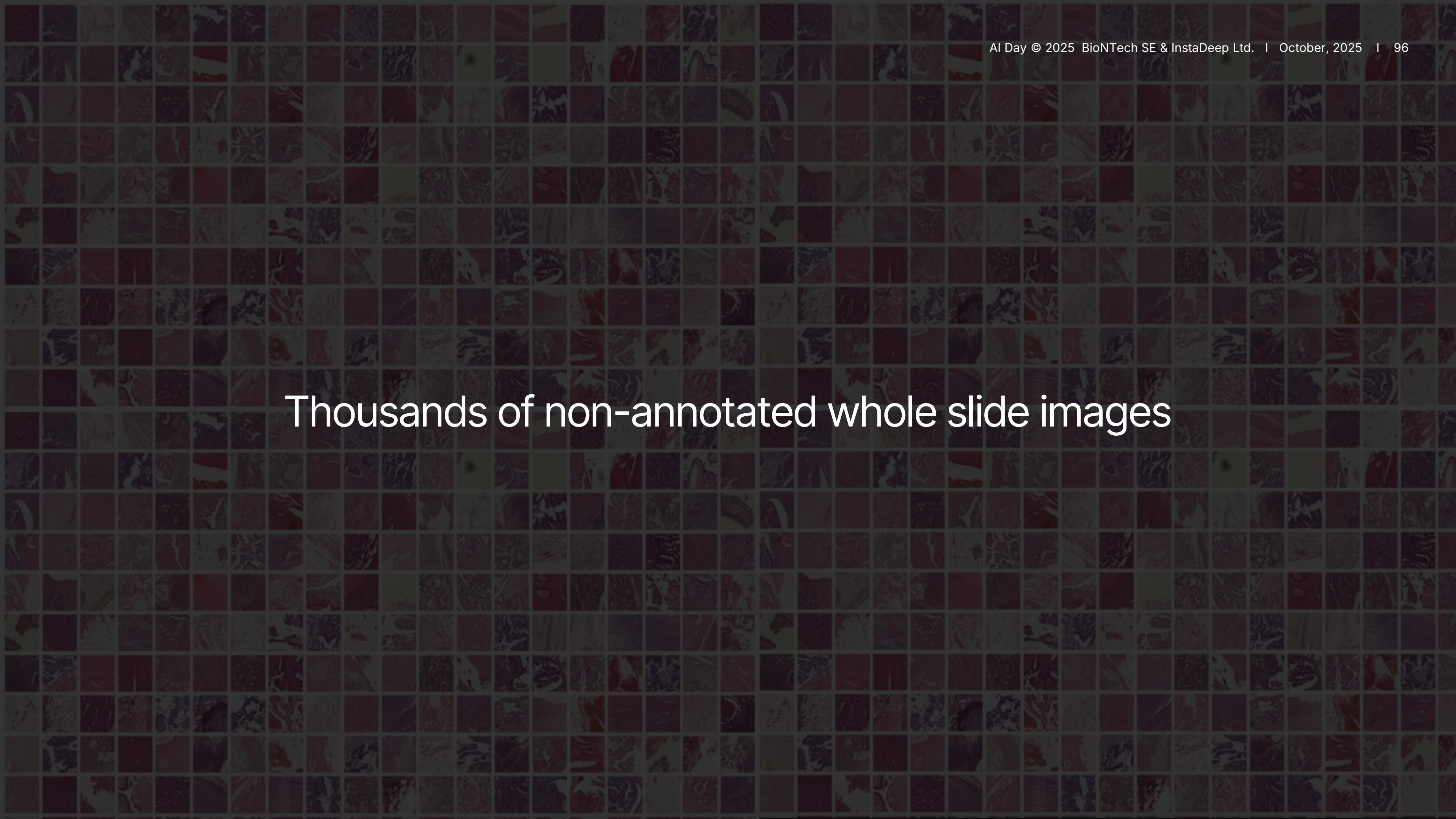


"*Introducing the next generation of InstaNovo models*", https://instanovo.ai/introducing-the-next-generation-of-instanovo-models/

# AI-Assisted tissue annotation tool (last year)

Increased the efficiency of pathologists fivefold (5x) compared to manual annotation.

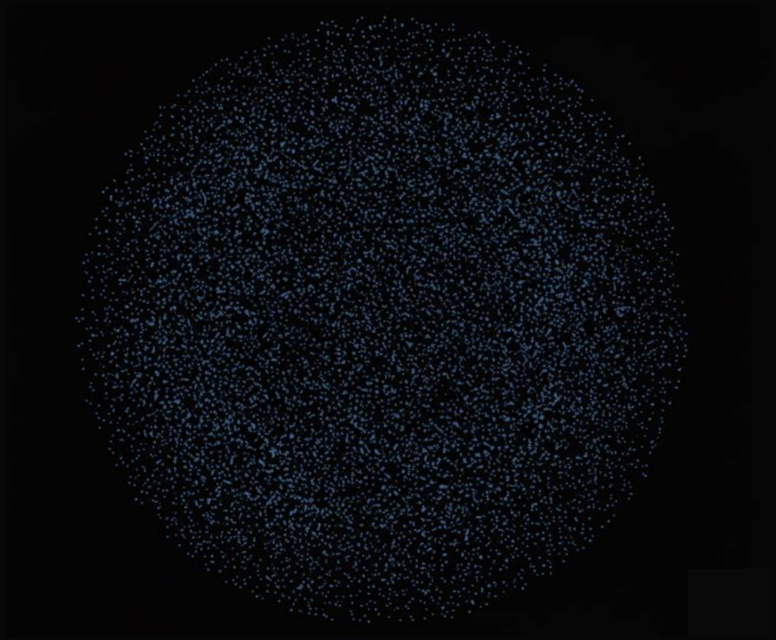
# 5× faster pathologists — but still not enough



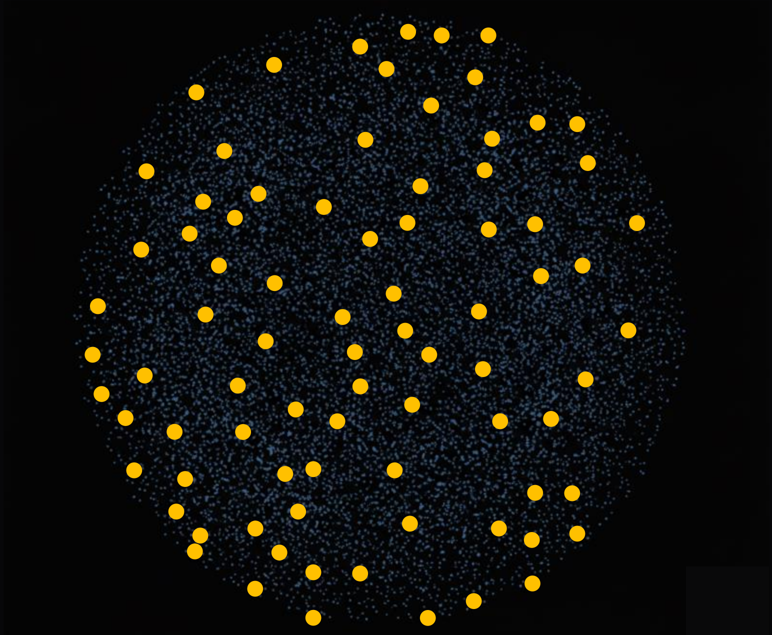
# Thousands of non-annotated whole slide images

# How can we reduce the pathologists’ annotation efforts while ensuring optimal model performance?

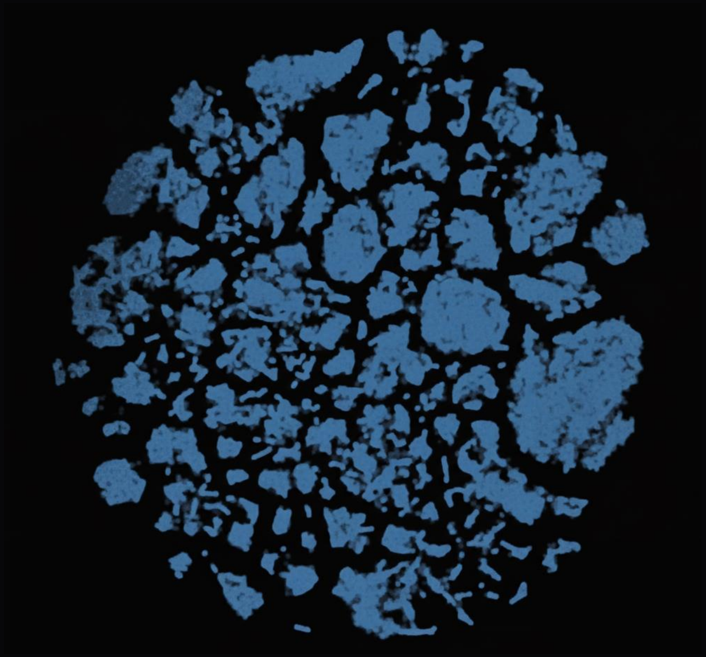
# Random data points selection



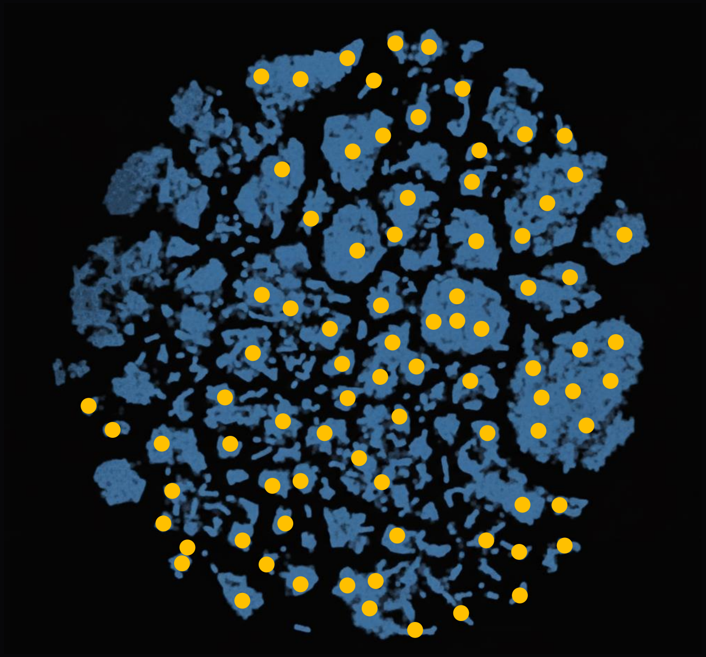
# Random Data Points Selection



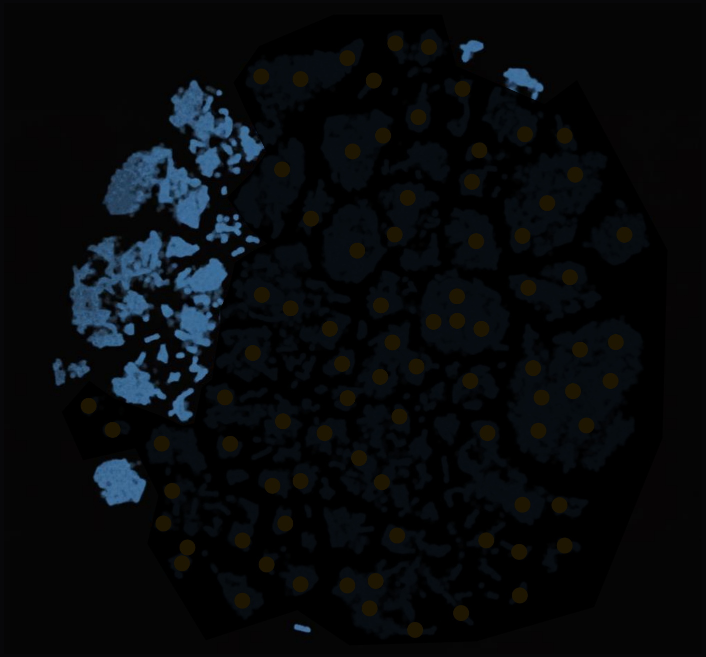
# Use Foundation Models to Cluster the Data by Patterns



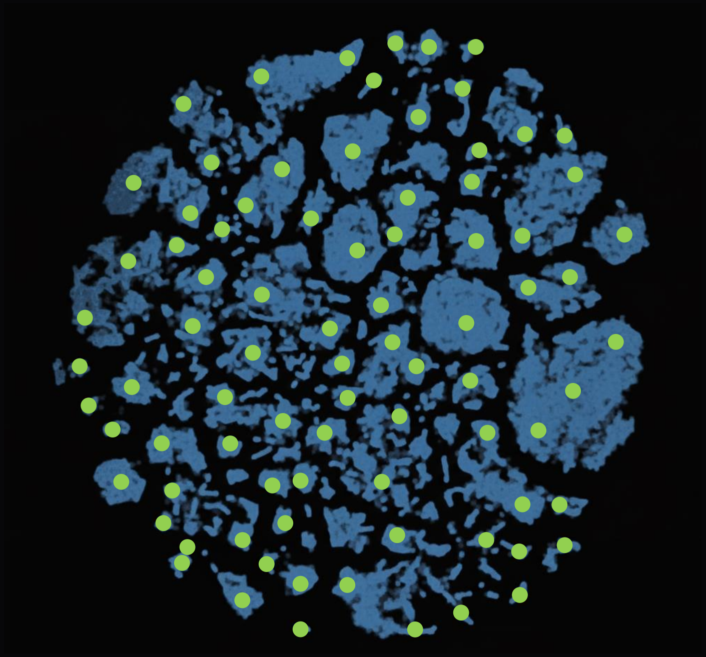
# Use Foundation Models to Cluster the Data by Patterns



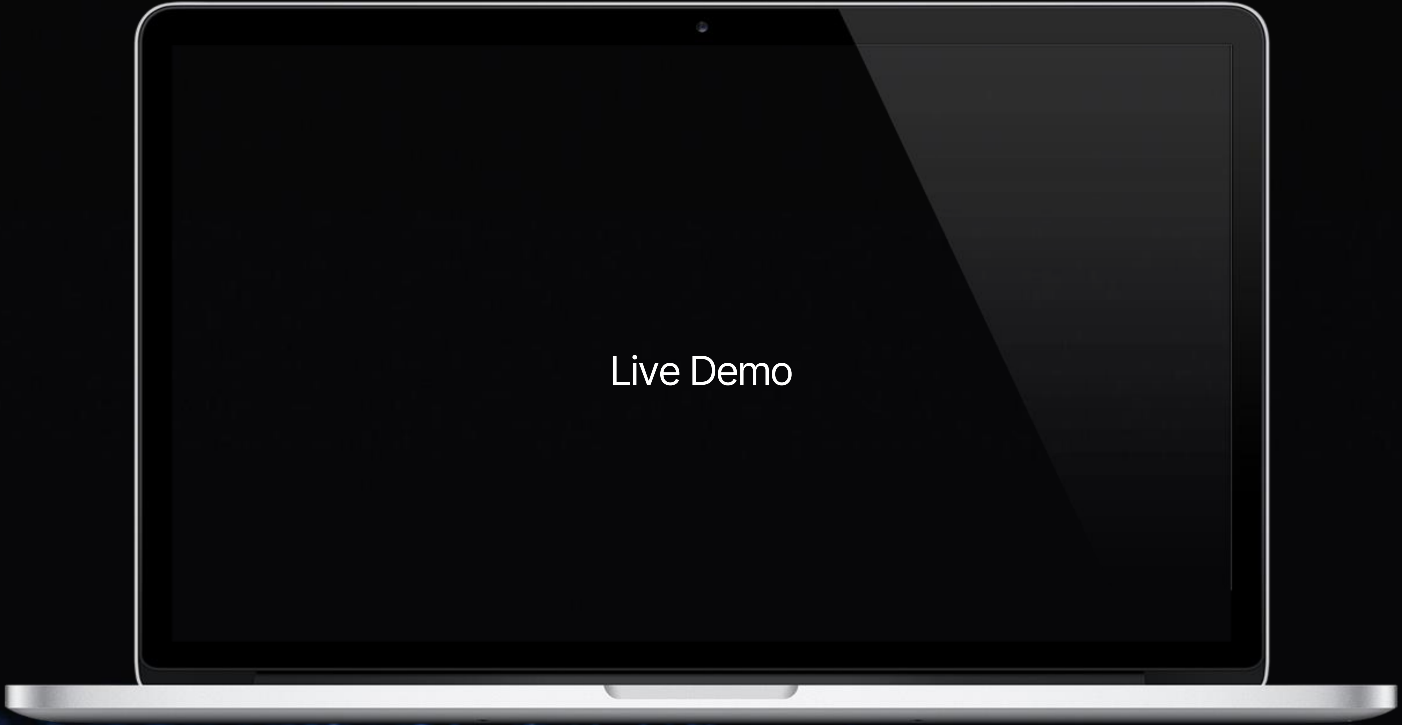
# Use Foundation Models to Cluster the Data by Patterns



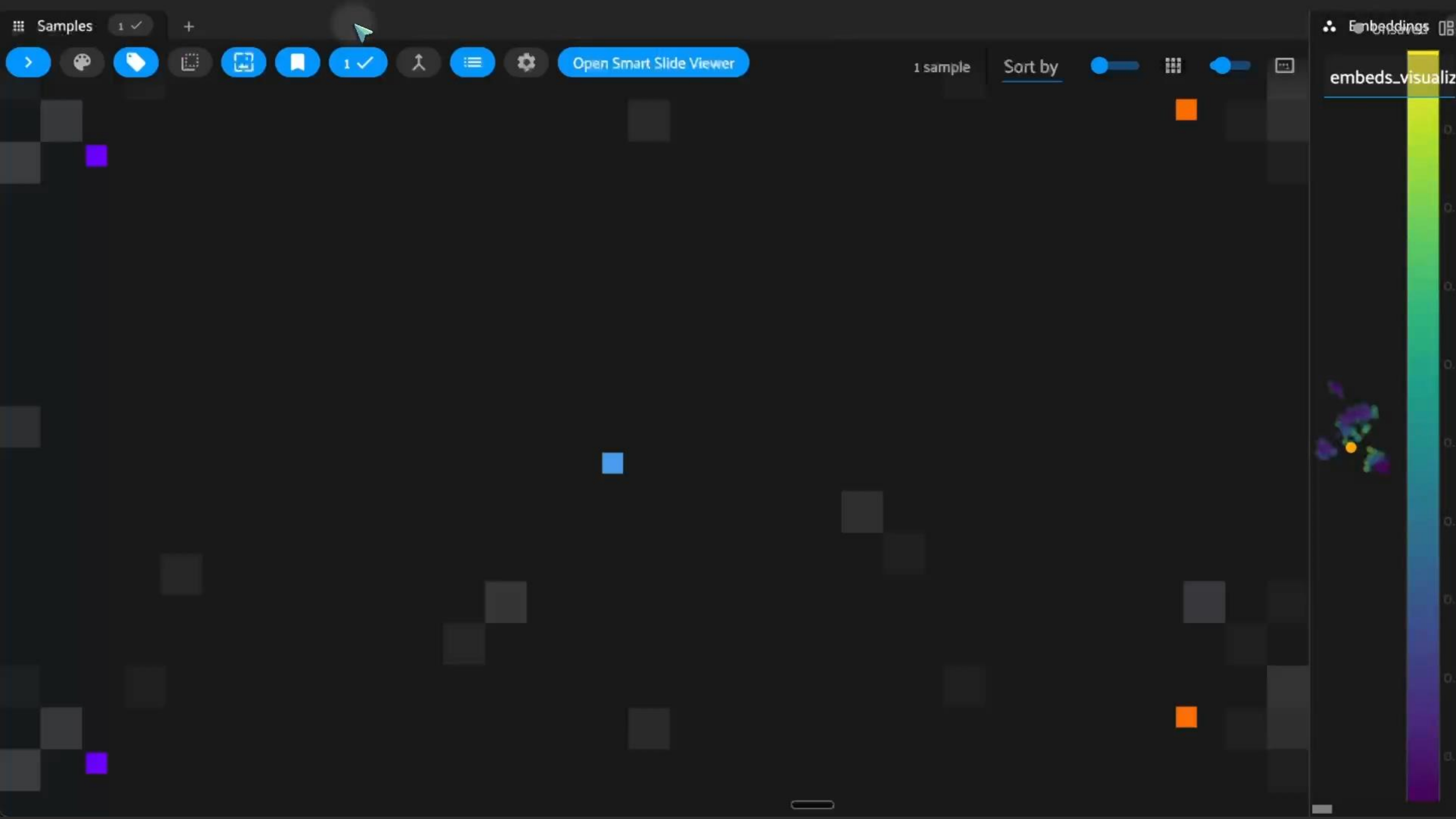
# Use foundation models + smart data points selection



# We developed a tool to explore, understand, and work with our histology data at scale.



# Live Demo


Samples
1 ✓
+
1 ✓
Open Smart Slide Viewer
1 sample
Sort by
Embeddings
embeds_visualiz

# Applications

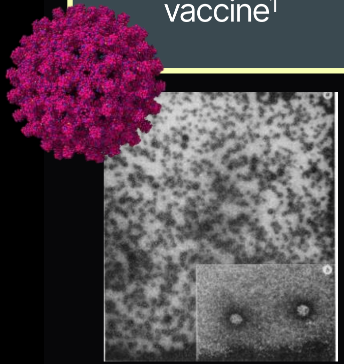
# Nanoparticle design

Cheng Zhang
Research Engineer
InstaDeep

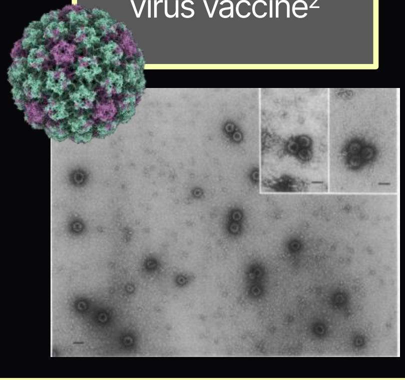
Lexi Walls
Senior Scientist II
BioNTech

# High valency nanoparticle vaccines yield strong antibody responses towards tough infectious disease targets

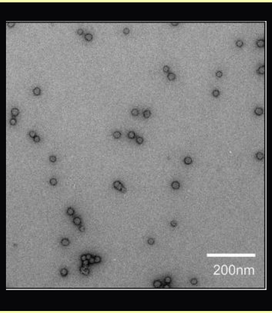
Hepatitis B vaccine[1]



Human papilloma virus vaccine[2]



Malaria vaccine[3]

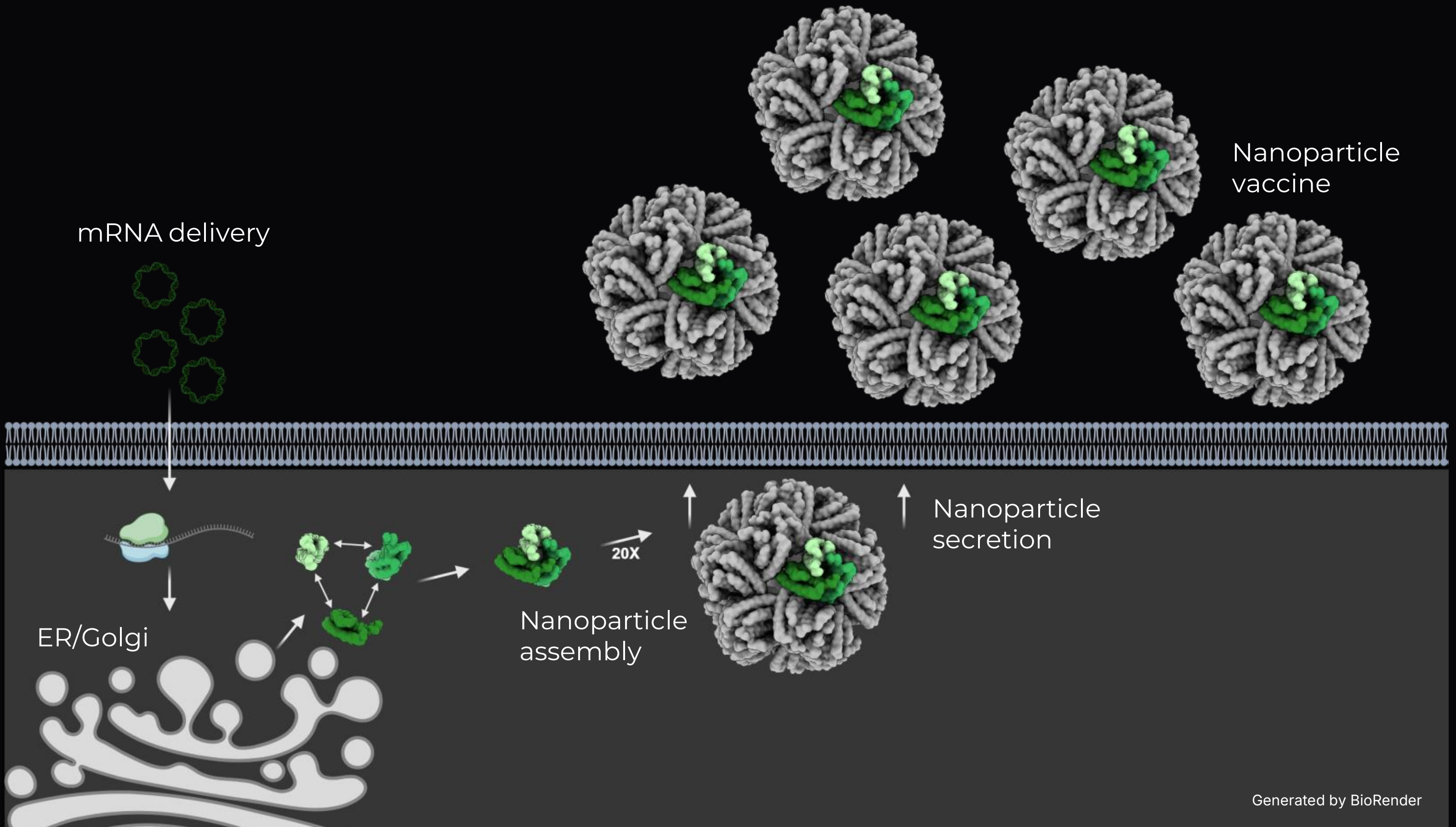


Nanoparticle vaccines have a crown of repeating antigen on a scaffold
They yield improved immune responses compared to solitary antigens
All nanoparticle vaccines in humans are protein based

**Goal:** Leverage AI to build nanoparticles suited to harness the power of mRNA vaccines

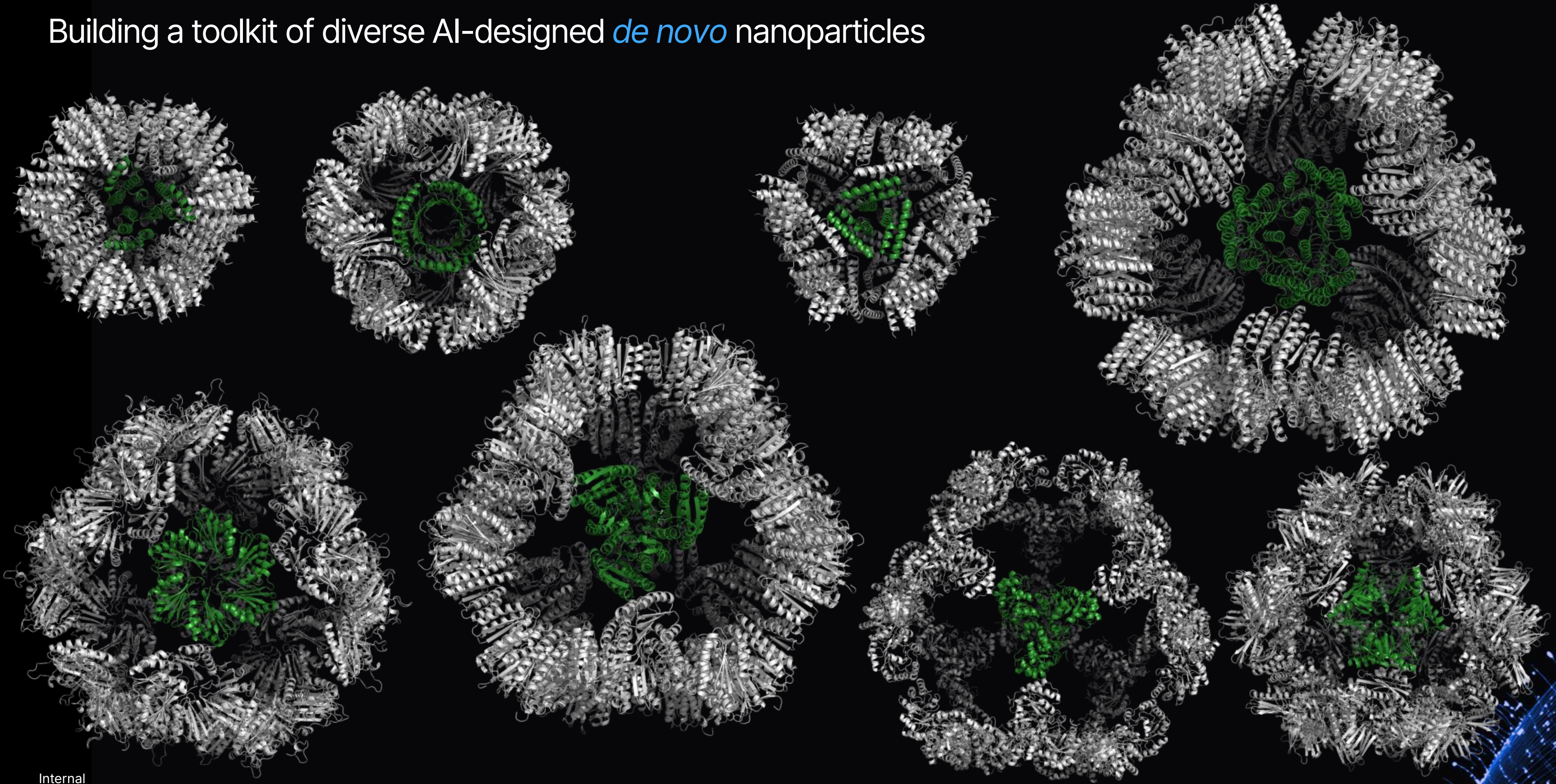
[1] Valenzuela et al. Nature. 1982.
[2] Kirnbauer et al. Proc. Natl. Acad. Sci. 1992.
[3] Collins et al. Sci Rep. 2017.

# Goal: mRNA launched nanoparticle vaccines

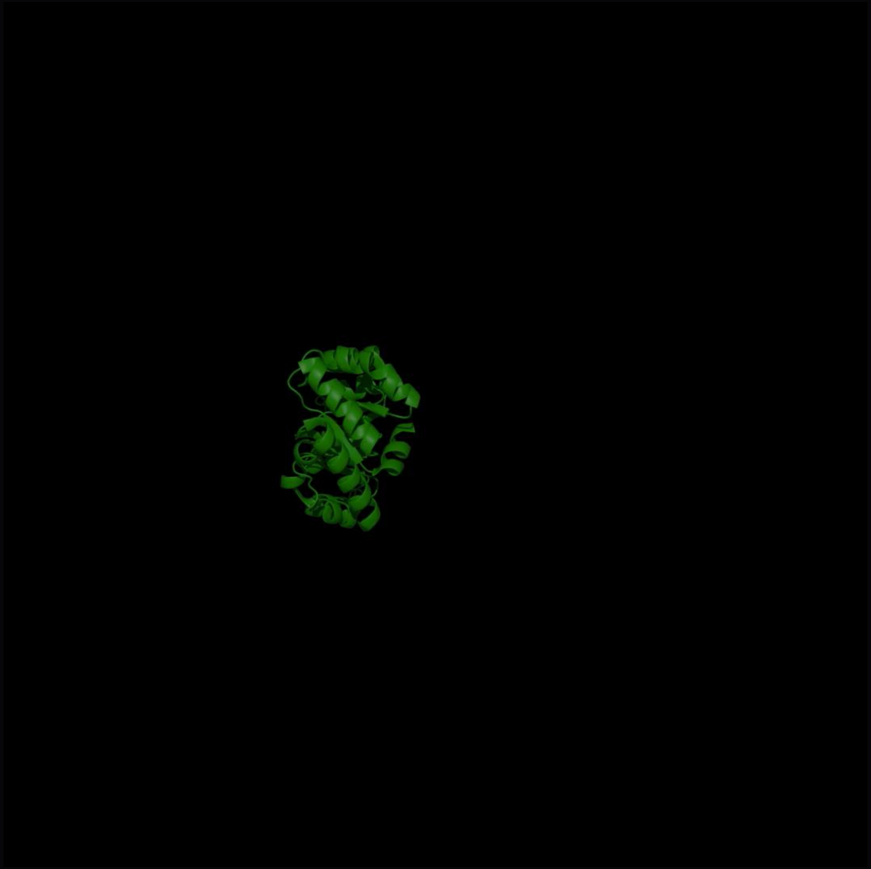


Generated by BioRender

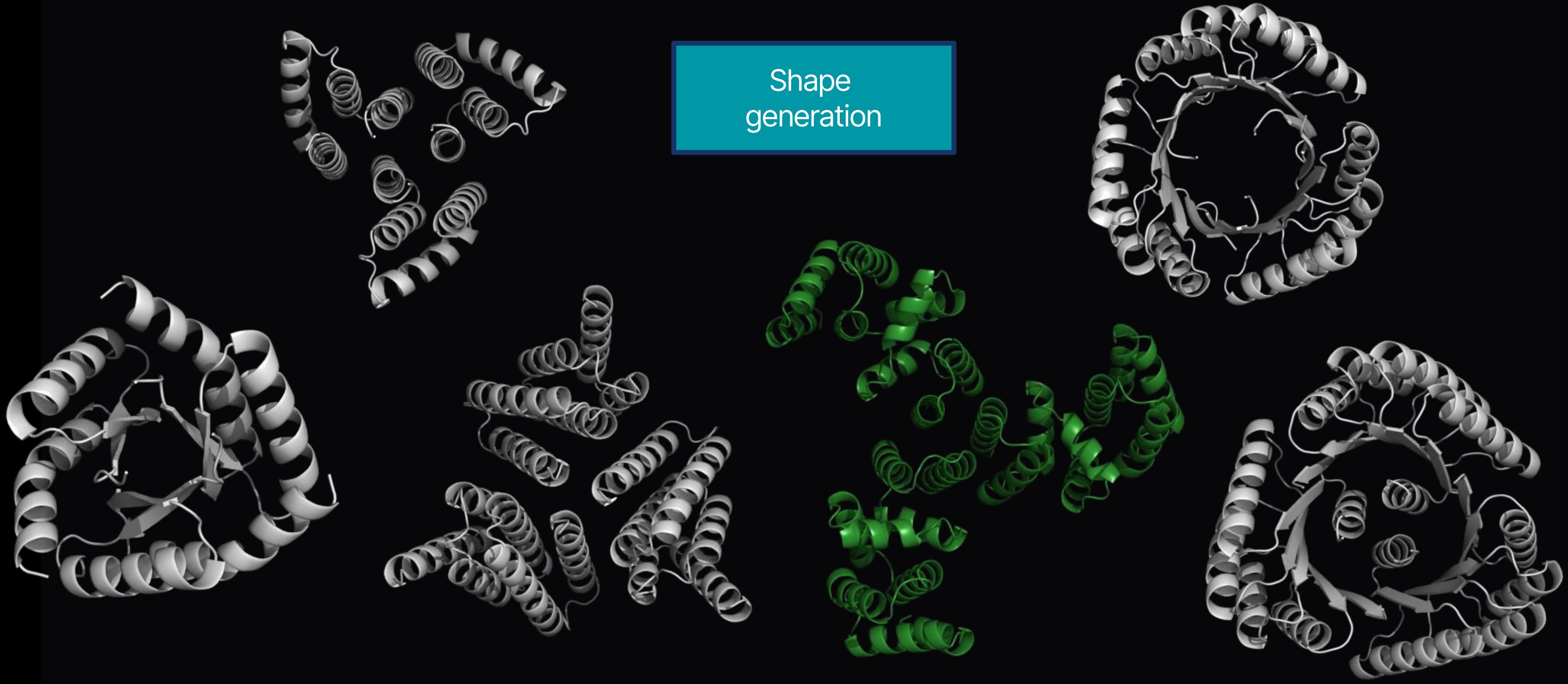
# Building a toolkit of diverse AI-designed *de novo* nanoparticles



# Building a nanoparticle piece by piece



# Utilizing AI protein design to build the nanoparticle components

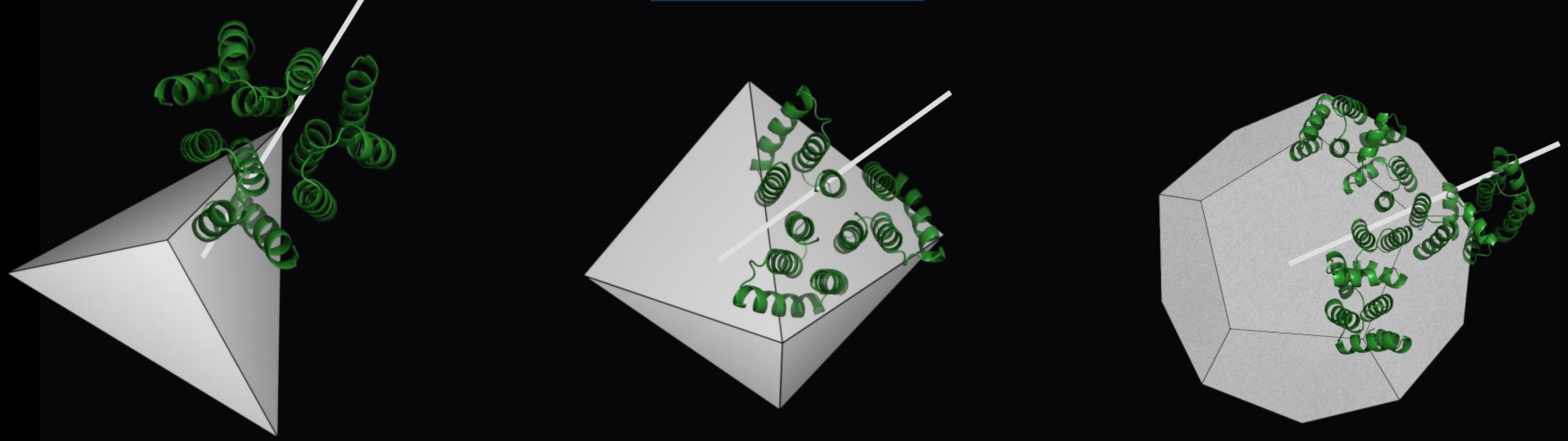


Shape generation

Generate thousands of de novo trimer shapes to enhance diversity of building blocks

# Assembling the nanoparticle building blocks into desired shapes

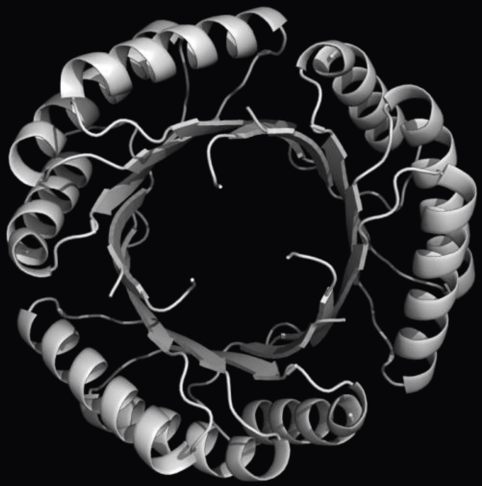
Symmetric assembling



Assembled to thousands of symmetric shapes

# Designing amino acid sequences to form the protein shape

Sequence design



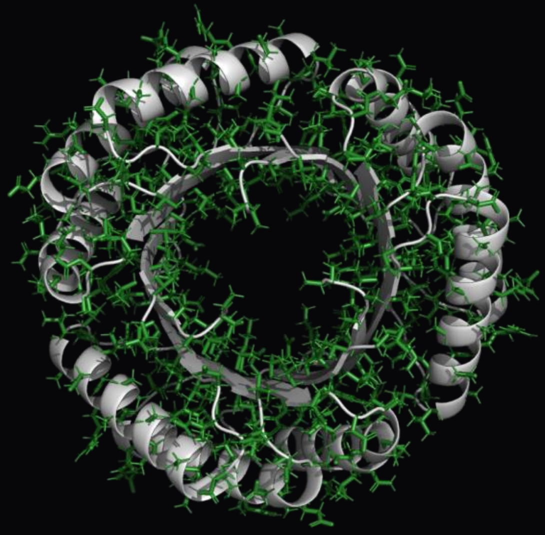
NLGVTFKWS...
VDEVTATQTH

→

Hundreds of sequences per particle

→

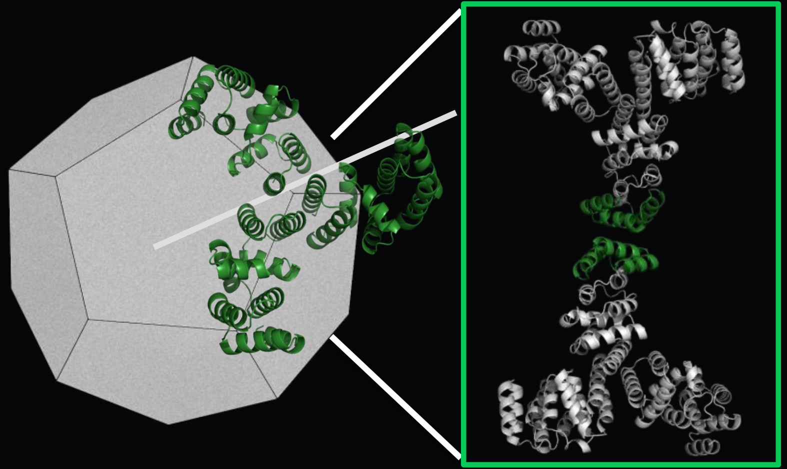
SPRHTLALR...
ATMKESVAE



Generate hundreds of thousands sequences to match the desired shapes and assemblies

# Computationally rank and filter the nanoparticle models

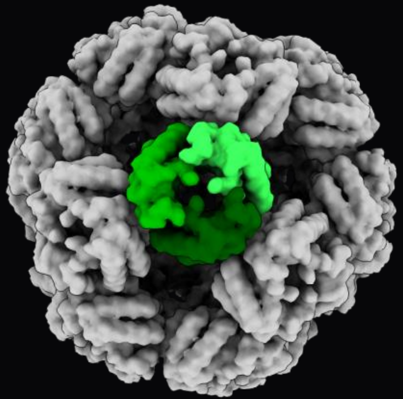
DeepChain
Folding Studio



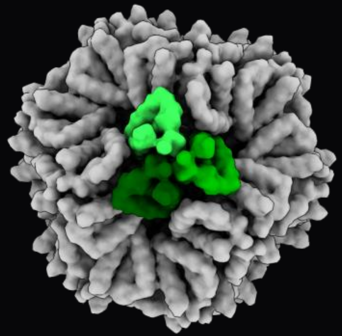
Filter and enrich to tens - hundreds of high-quality designs prior to laboratory testing

# *In vitro*: confirming nanoparticle design and assembly

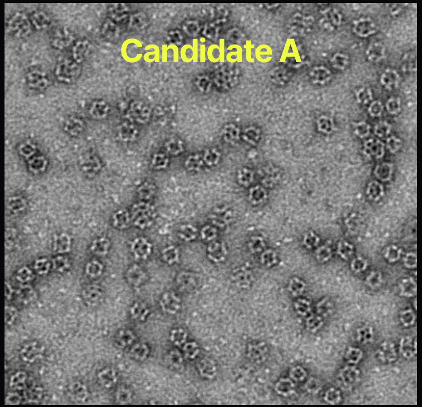
Computationally designed nanoparticle model

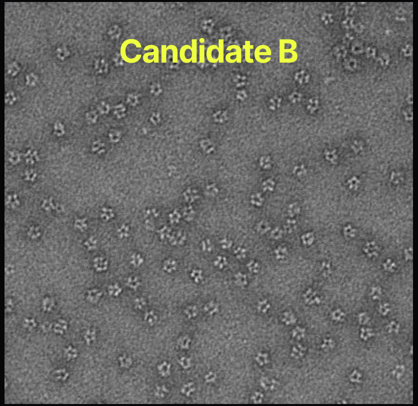


Wet-lab experiments



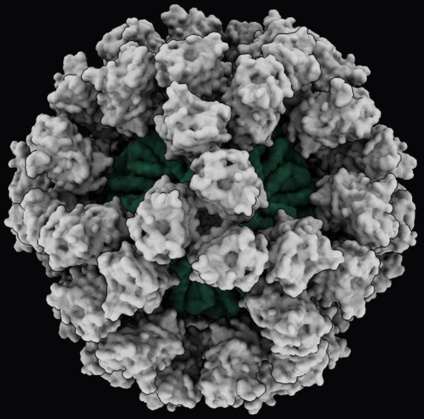
Negative stain electron micrographs confirming nanoparticle assembly



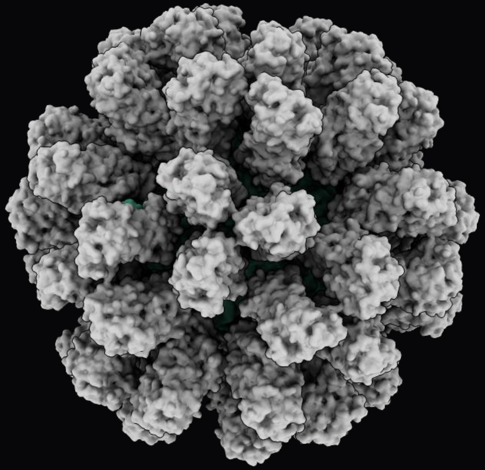


# *In vitro*: showcasing nanoparticles can display vaccine antigens

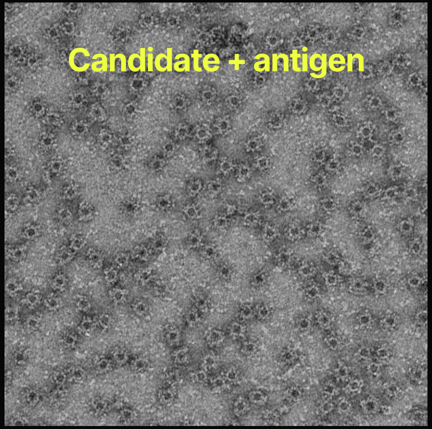
Computationally designed nanoparticle + antigen model

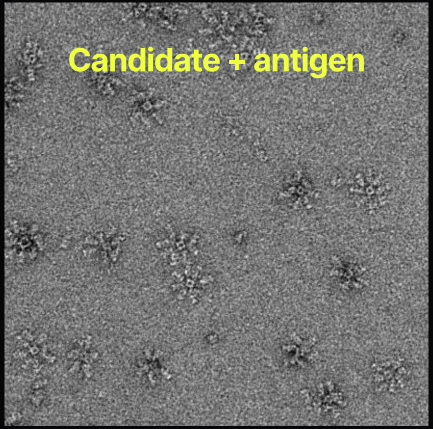


Wet-lab experiments



Negative stain electron micrographs confirming nanoparticle displays antigen





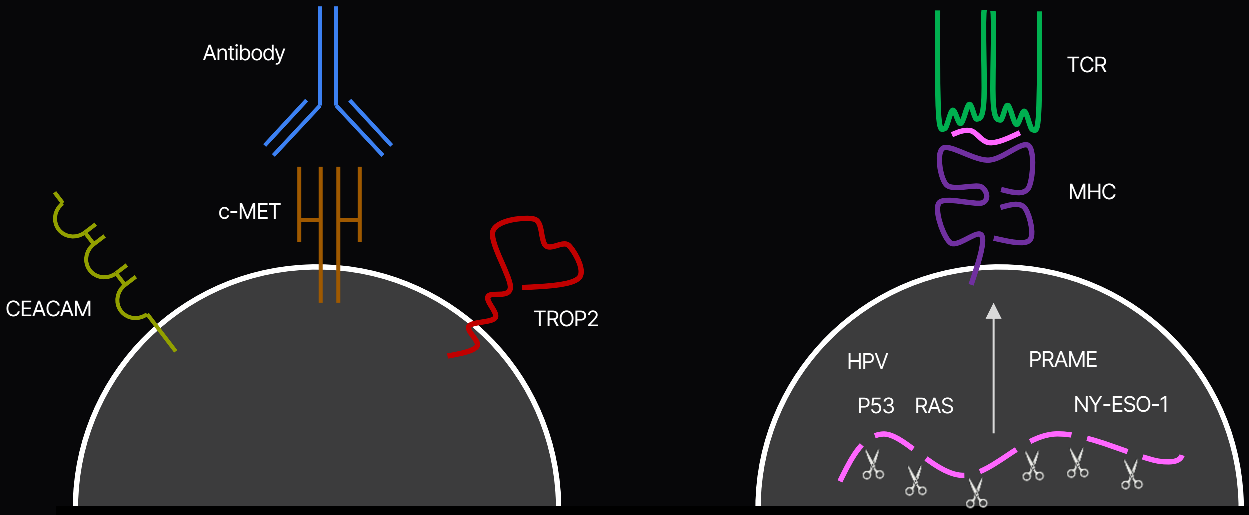
Can you put this in as my head shot? Currently the slide shows John.
MY title is senior director, computational biology

# TCR affinity enhancement

Antoine Delaunay
Senior Research Engineer
InstaDeep

Michael Rooney
Senior Director Comp Biology
BioNTech

# T cell receptors (TCRs) can access highly tumor-specific cancer targets

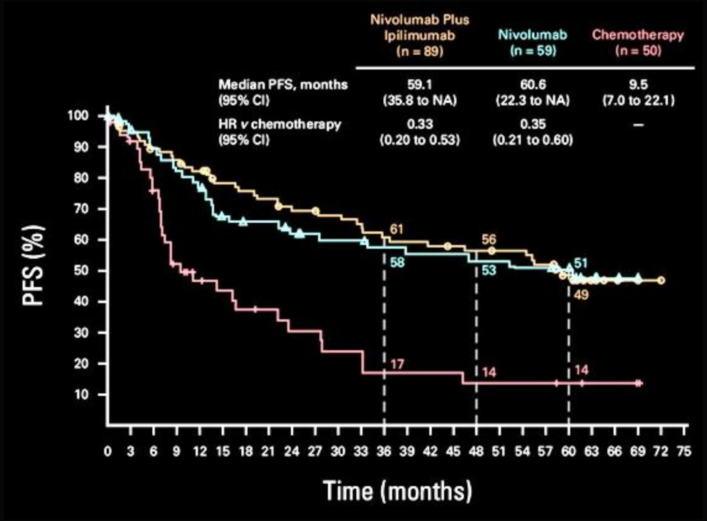


Extracellular targets
*Modest tumor specificity*

Intracellular targets
*High tumor specificity*

# T cells can achieve durable responses

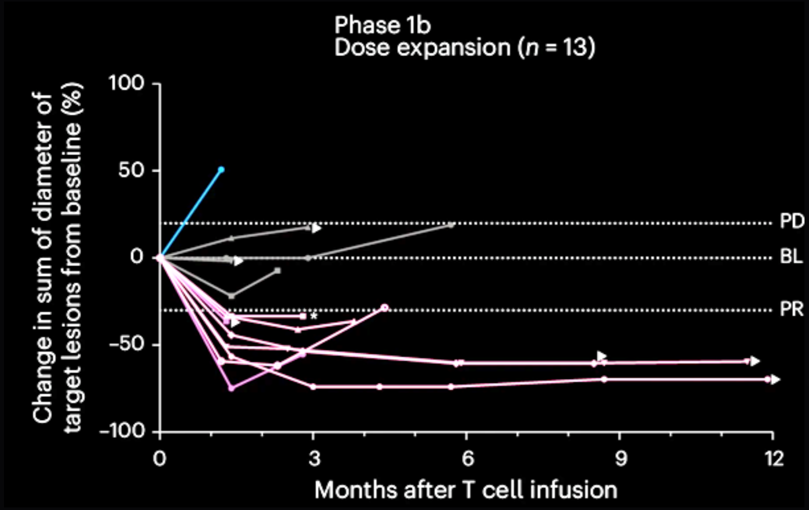
## Checkpoint blockade in NSCLC



| | Nivolumab Plus Ipilimumab (n = 89) | Nivolumab (n = 59) | Chemotherapy (n = 50) |
|---|---|---|---|
| Median PFS, months (95% CI) | 59.1 (35.8 to NA) | 60.6 (22.3 to NA) | 9.5 (7.0 to 22.1) |
| HR v chemotherapy (95% CI) | 0.33 (0.20 to 0.53) | 0.35 (0.21 to 0.60) | – |

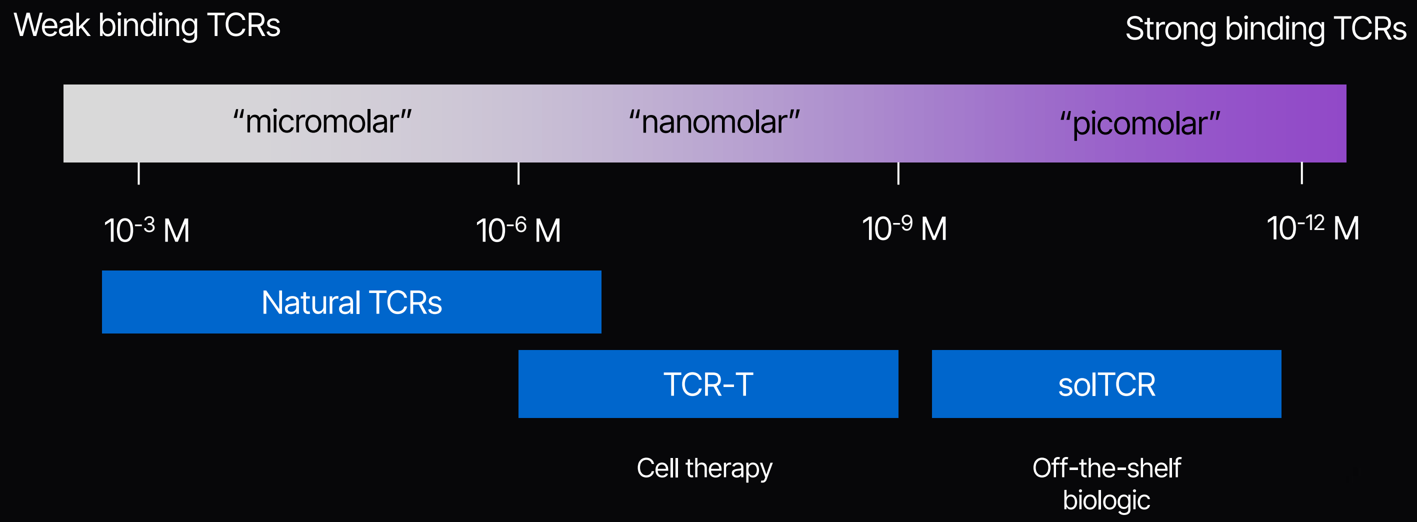
Brahmer, JCO, 2023.

## PRAME-directed TCR-T



Wermke, Nature Medicine, 2025.

# Affinity enhancement is required to unlock the full potential of T cell receptors (TCRs)



Weak binding TCRs

Strong binding TCRs

“micromolar” “nanomolar” “picomolar”

$10^{-3}$ M $10^{-6}$ M $10^{-9}$ M $10^{-12}$ M

Natural TCRs

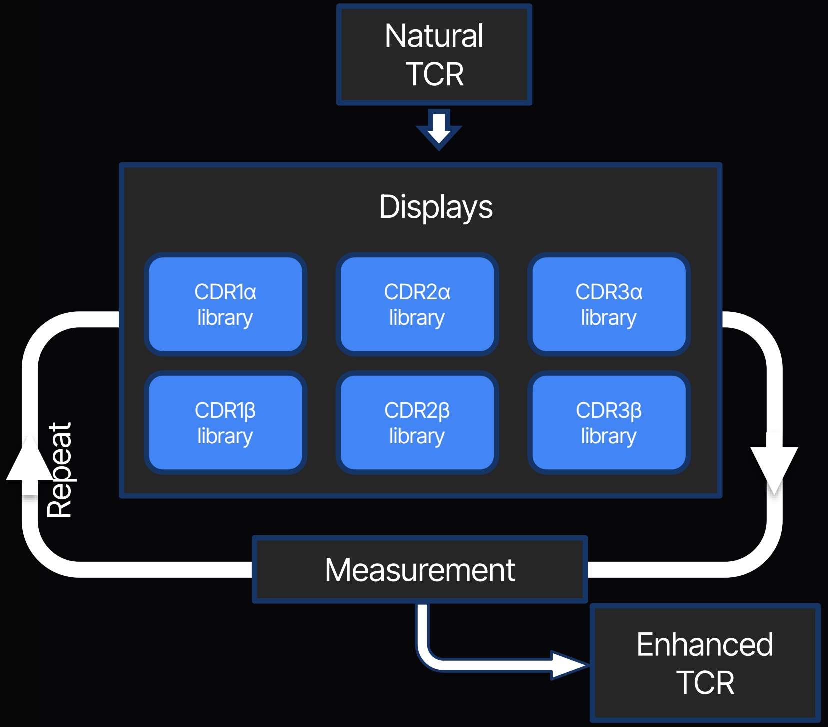
TCR-T

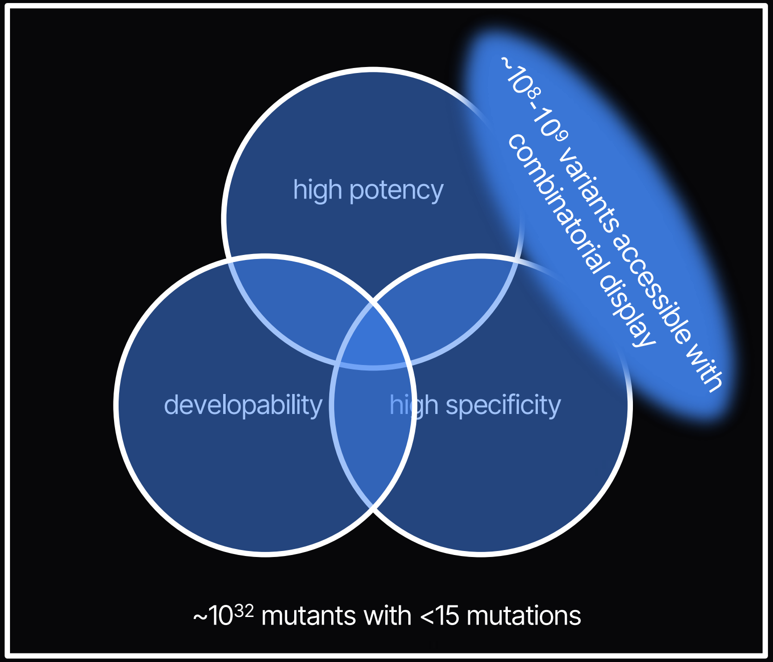
Cell therapy

solTCR

Off-the-shelf biologic

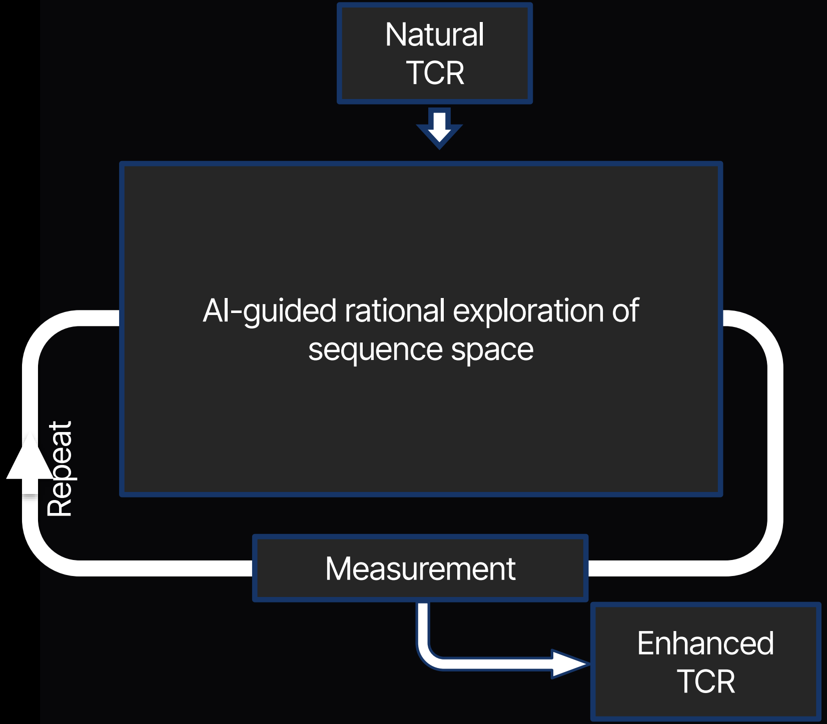
# Conventional display-based affinity enhancement is labor- intensive but explores tiny sliver of TCR sequence space

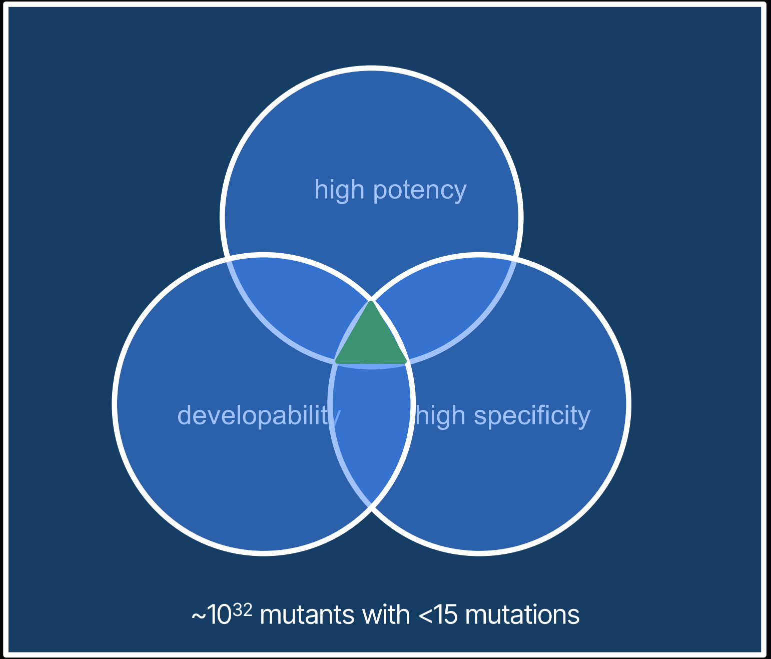




~$10^{32}$ mutants with <15 mutations

# AI-guided exploration of TCR sequence space enables efficient discovery of optimized variants

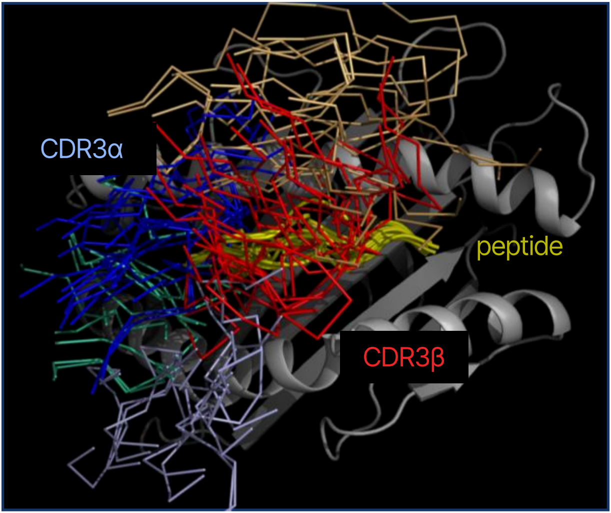




~$10^{32}$ mutants with <15 mutations

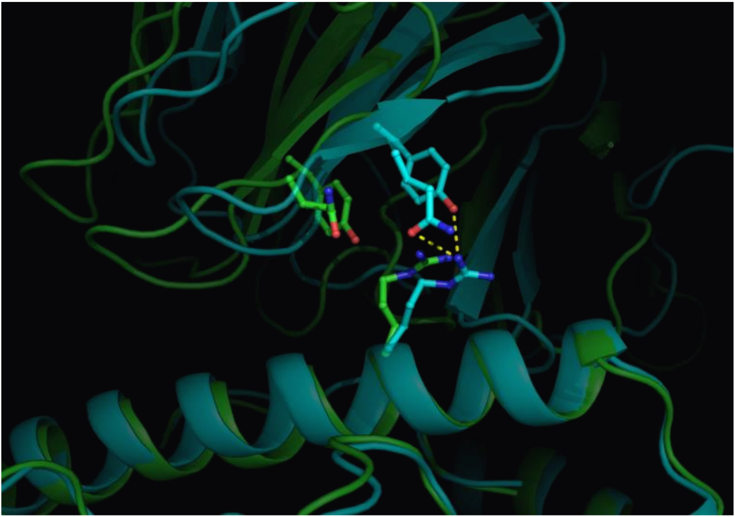
# Learning the rules of TCR optimization is hard due to high structural diversity of TCRpMHC interactions

## Overall TCR:pMHC docking is similar, but exact CDR loop positions are highly diverse



Twelve TCRpMHC structures superimposed by MHC
(PDB ID: 1ao7, 1mi5, 2ak4, 2nx5, 2ypl, 3dxa, 3ffc, 3h9s, 3vxm, 4g8g, 4jrx, 4mji)

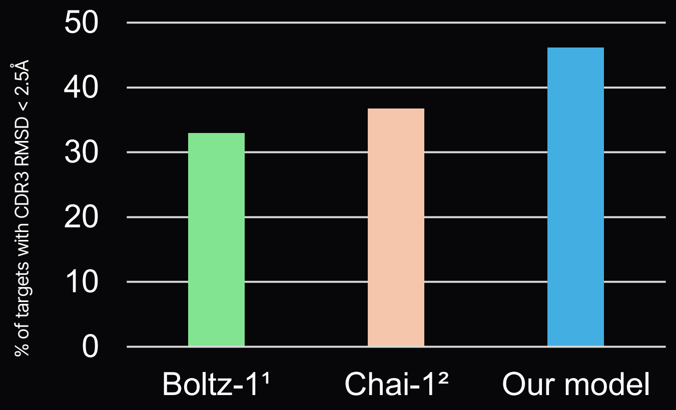
## Residue environment determines optimal substitutions but varies from TCR to TCR



Divergent germline CDR2β-MHC interactions in two structures that share both V-genes and MHC allele
(PDB ID: 5nht, 6vm9)

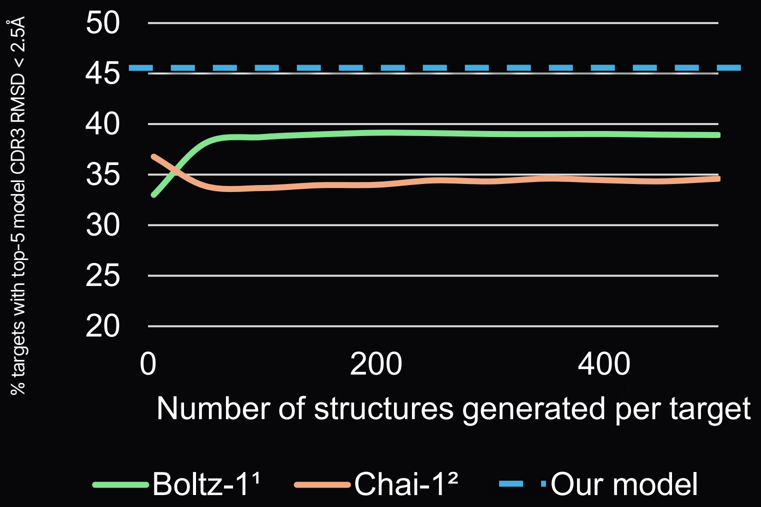
# Our model outperforms state-of-the-art in TCR–pMHC structure prediction

## Performance benchmark on test targets (CDR3 RMSD < 2.5Å )



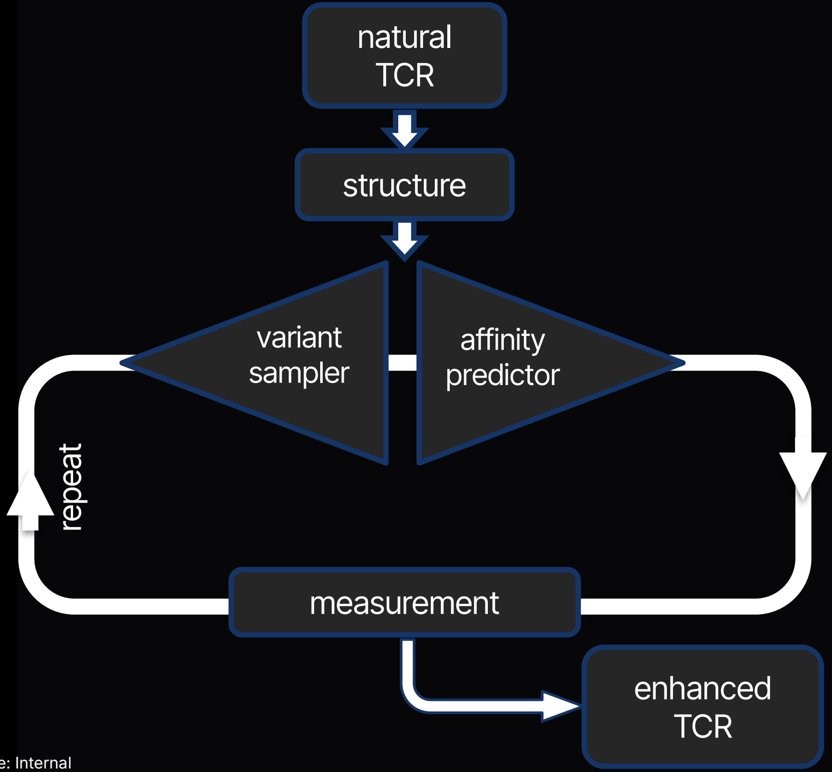
For each model, fraction of targets where at least one of 5 generated structures achieved CDR3 RMSD < 2.5Å. Test set contains only unseen targets.

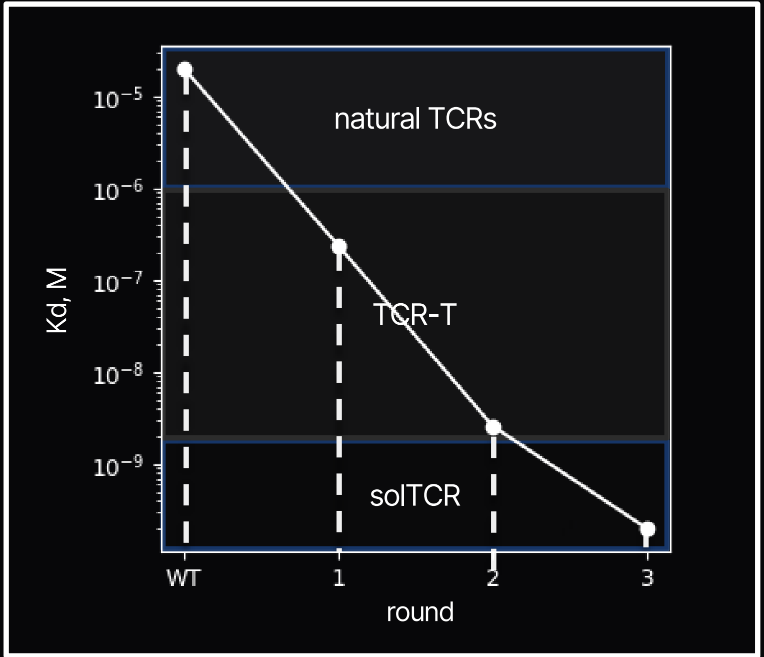
## Our method outperforms sampling models that quickly saturate



[1] Wohlwend et al., Boltz-1: Democratizing Biomolecular Interaction Modeling, *bioRxiv*, 2025.
[2] Chai Discovery et al., Chai-1: Decoding the molecular interactions of life, *bioRxiv*, 2024.

# Our AI pipeline achieves an average 50,000-fold TCR binding enhancement increase over WT, in three rounds or less, on the four considered targets. We repeatedly reach picomolar affinity.



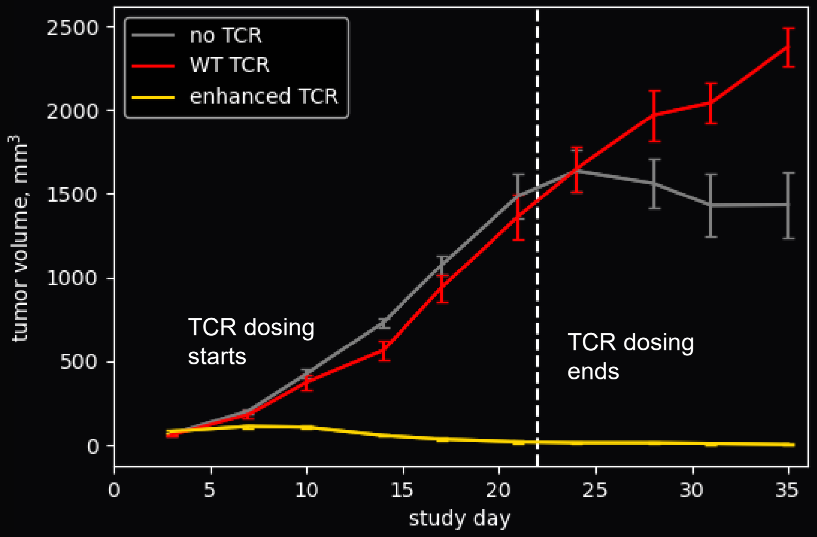


Example of a TCR affinity improvement of more than a 100,000-fold in 3 rounds

Source: Internal

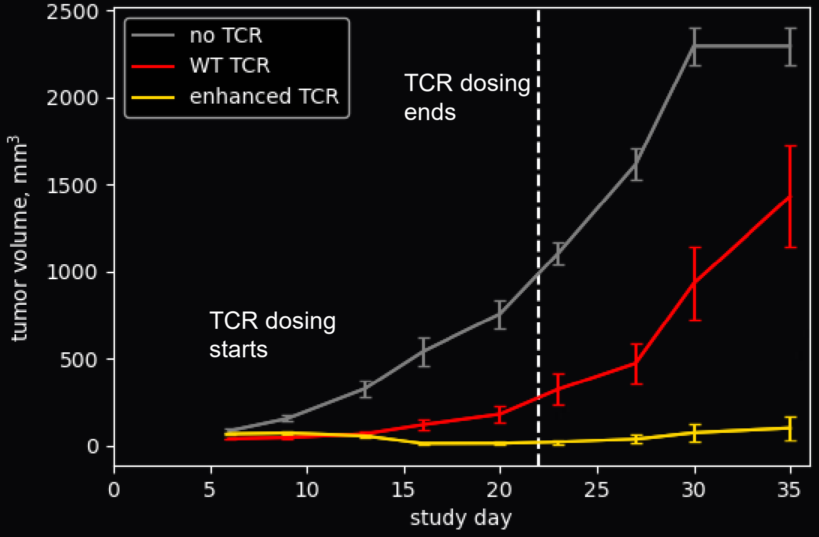
# Affinity-enhanced TCRs lead to strong and durable *in vivo* tumor control in a pre-clinical model

## pHLA tumor target 1

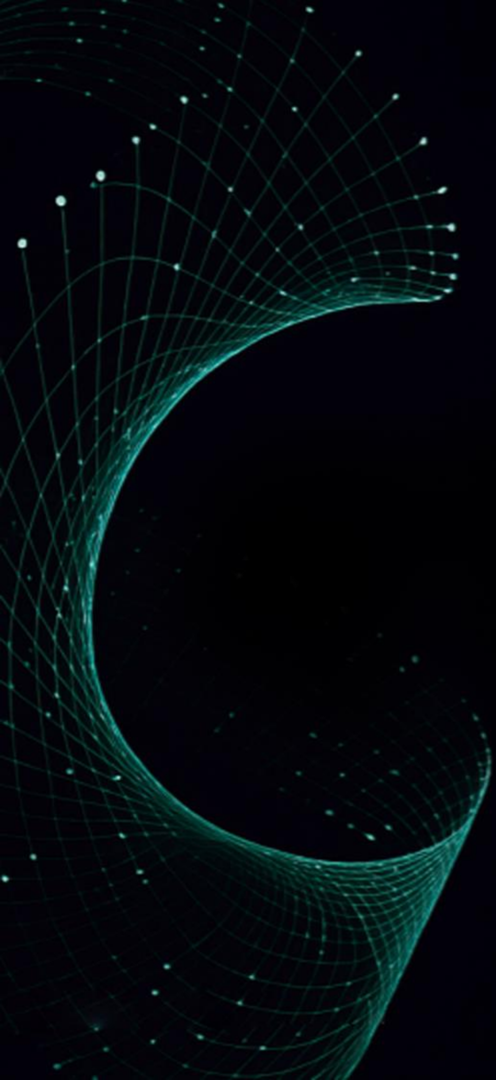


Source: Internal

## pHLA tumor target 2



Source: Internal



BIONTECH | InstaDeep

# AI Day

## Thank you for your time
## Next: Q&A

**October 2025**