VantAI

ILUMINATE THE MOLECULAR WORLD

Neo-1 | 2025

Computational approaches for small molecule drug discovery

Introduction: computers and drugs

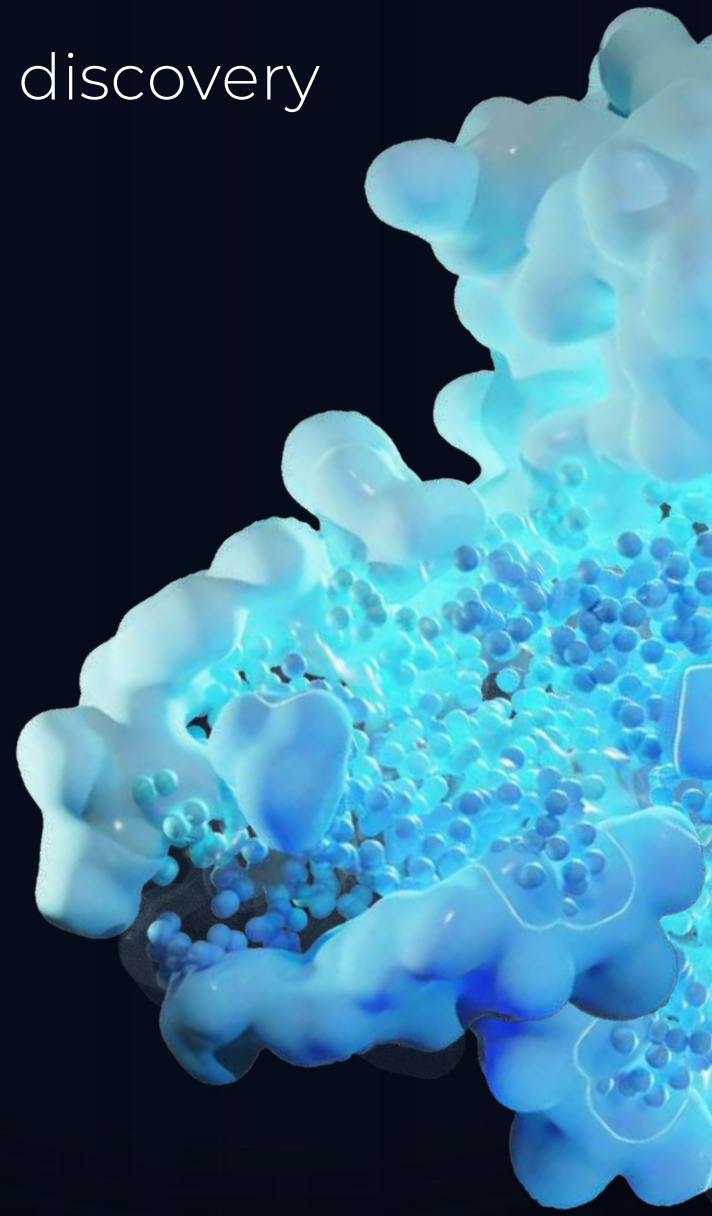
- Introduction
- The problem: the art and science of making drugs
- A quick history of computational drug discovery

The “classic” era: task-specific tools and models

- Single task focused methods
- Breakthrough in protein structure prediction: the road to AF2 and beyond
- De-novo molecular generators

The new era: foundation models trained on black box data – Neo and beyond

- Neo-1: unifying all-atom structure prediction and de-novo generation for the first time
- The promise of black-box data & NeoLink



VANTAI

Who are we?

Luca Naef
Co-Founder &
CTO

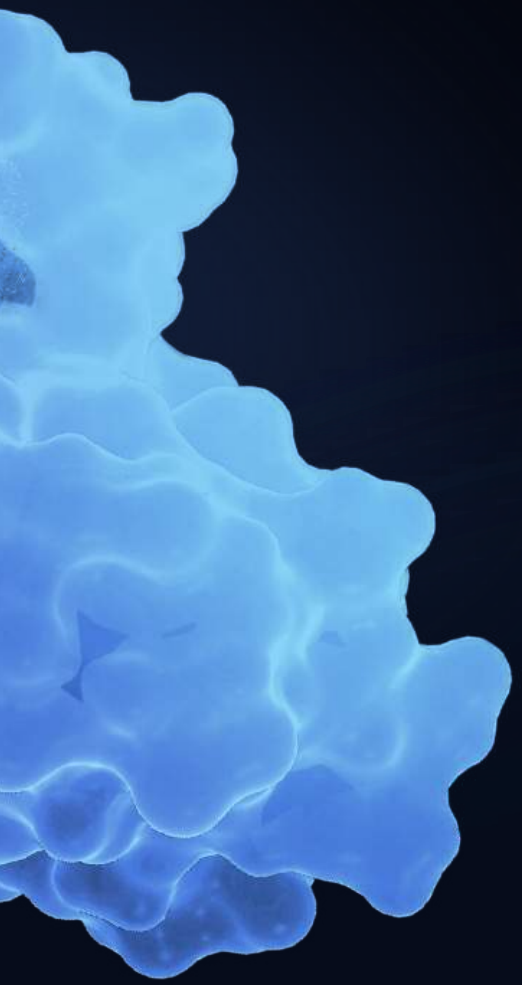
- BSc/MSc ETH Zurich – Molecular Bio & Deep Learning
- Research in Stanford, UNSW, TokyoTech
- Software Engineer & first tech startup during BSc
- Diverse roles in Biotech – Regeneus (AUS), CJ Partners (JPN)
- QuantumBlack & McK – AI in Drug Discovery across Fortune 100/500
- Co-founded VantAI in 2019

Vladas Oleinikovas
Director of Comp Chem

- BA/MSci Cambridge – Nat. Sci. / Chem.
- PhD UCL - Chemistry
- Senior Scientist at UCB Pharma
- Acting Head of CADD at Monte Rosa Tx
- Co-inventor of clinical VAV1 degrader
- Joined VantAI in 2024

VANTAI

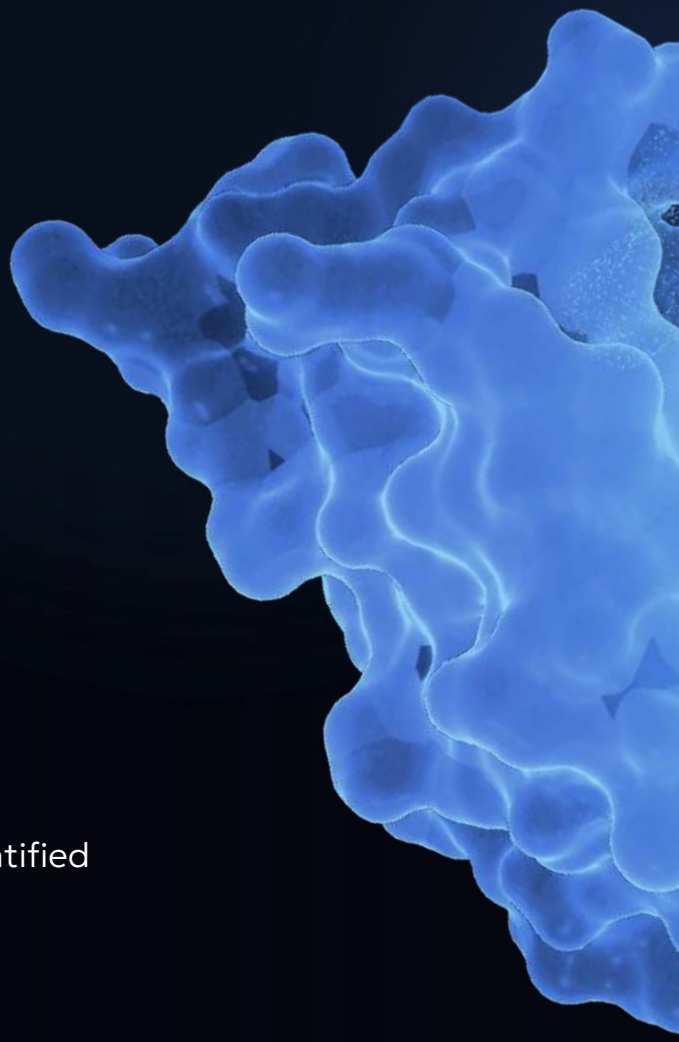
Finding new medicines is hard



~1-3B Cost to find a new drug[1]

90+% Chance of failure **after** entering human trials[2]

13+ yrs Typical development **after** cause of disease identified



1. Estimates vary – e.g. from 0.88B (Eastern Research Group, "Drug Development Final Report", Sept. 2024, for U.S. Department of Health and Human Services) to 2.6B (DiMasi et al., J Health Econ 2016)
2. Smietana et al., Nature Reviews Drug Discovery, 2016
3. Paul et al., Nature Reviews Drug Discovery, 2010

VANTAI

Disease

Tuberculosis

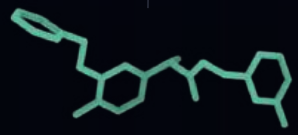
10+ years
++M USD

Target

PKS13

10+ years
++M USD

Drug



VANTAI

And has been getting harder

Moore's Law: Transistors

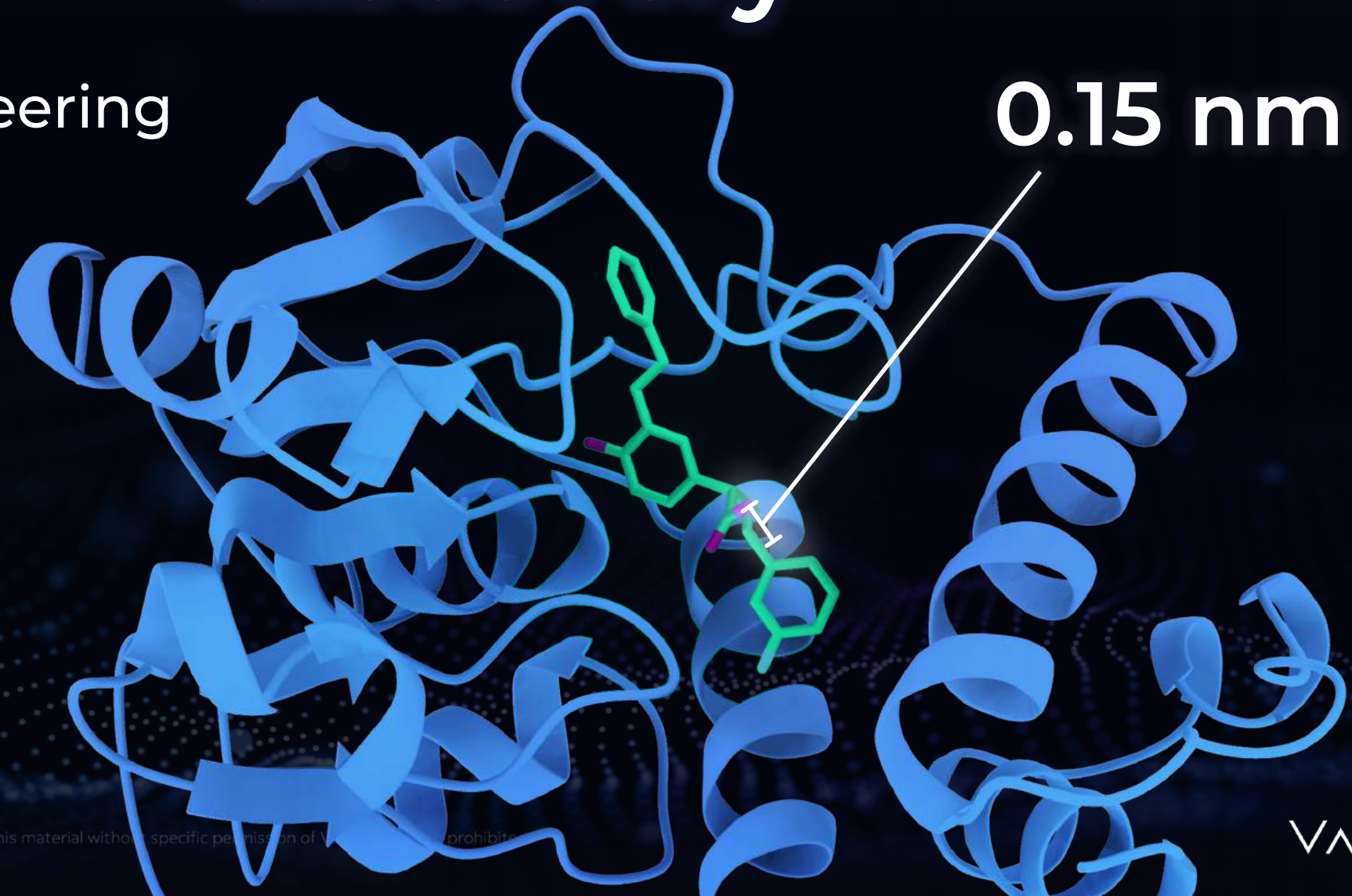
Eroom's Law: Medicines

1. Unite.ai, 2025
2. Scannel et al., Nature Reviews Drug Discovery, 2012

VANTAI

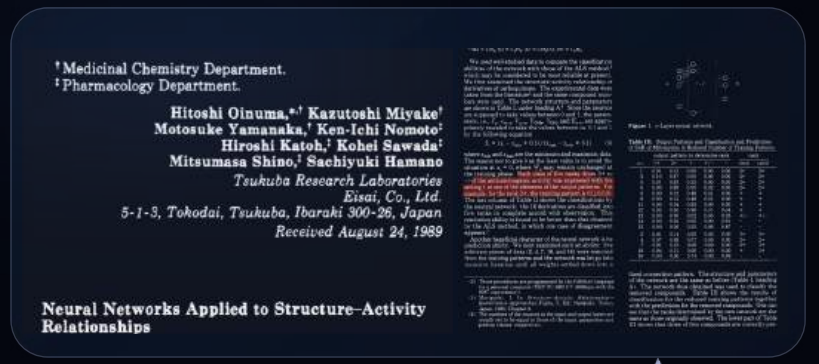
Goal: Rational Drug discovery

Precision nano-engineering of therapies



VANTAI

This is, by no means, a new idea!



† Medicinal Chemistry Department.
‡ Pharmacology Department.

Hitoshi Ojnuma,*,† Kazutoshi Miyake† Motosuke Yamanaka,† Ken-Ichi Nomoto‡ Hiroshi Katoh,‡ Kohei Sawada‡ Mitsumasa Shino,‡ Sachiyuki Hamano

Tsukuba Research Laboratories Eisai, Co., Ltd.
5-1-3, Tokodai, Tsukuba, Ibaraki 300-26, Japan
Received August 24, 1989

Neural Networks Applied to Structure–Activity Relationships

The computer program LUDI: A new method for the de novo design of enzyme inhibitors

Hans-Joachim Böhm

BASF AG, Central Research, D-6700 Ludwigshafen, Germany

Received 27 May 1991
Accepted 16 August 1991

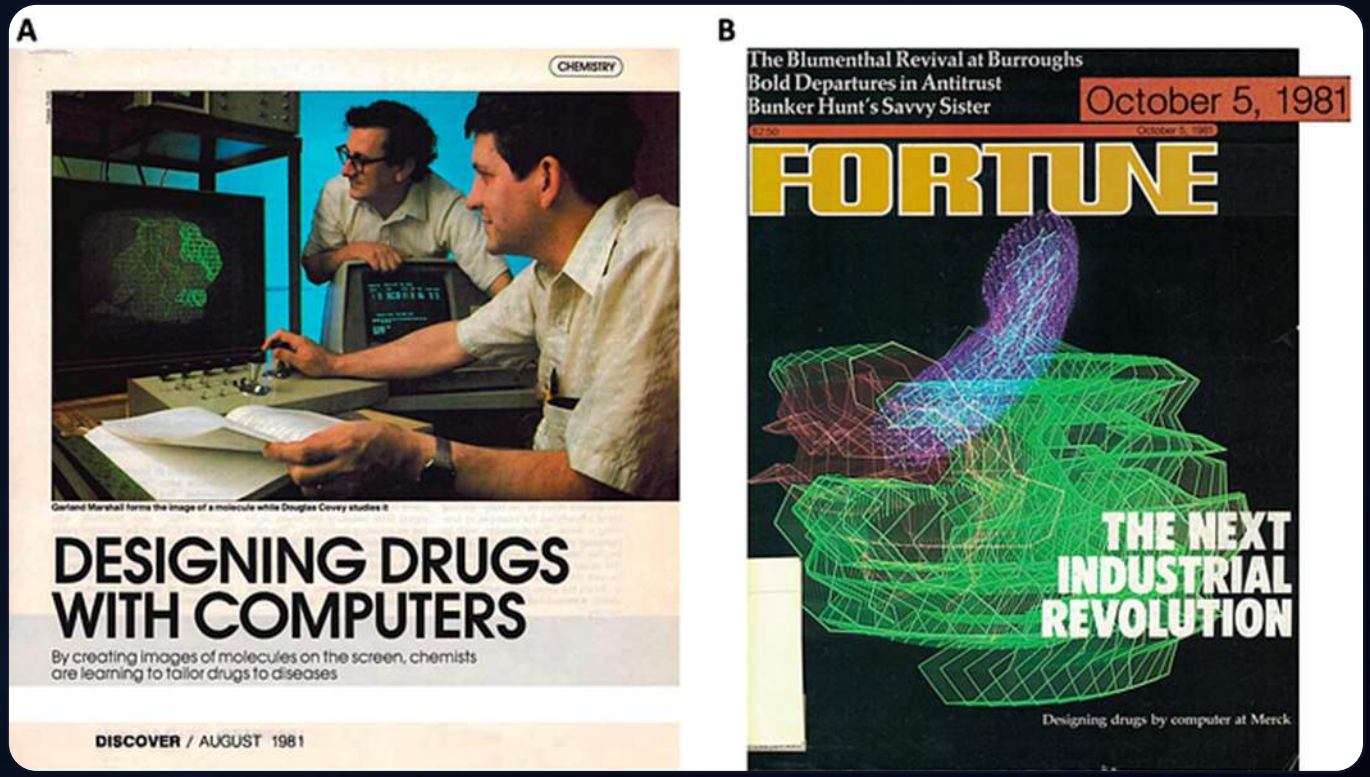
Key words: Enzymes; Enzyme inhibitors; Molecular modeling; Drug design; De novo design

64’	81’	82’	90’	92’	95’	2015	2020	2024
QSAR Models via regression: Corwin Hansch and Toshio Fujita	Captopril: first “rational” structure-based Drug approved in 1981	Early Docking tools: DOCK, Irwin D. Kuntz	NN for QSAR. Toshihisa Aoyama et al. (J. Med. Chem., 1990)	“De-novo” generative method via rules-based fragment assembly. Boehm, 1992, Journal of Computer-Aided Molecular Design	Computational modelling drives new class of anti-HIV drugs: saquinavir (approved 1995) indinavir (1996), ritonavir (1996), and nelfinavir (1997	AtomNet. Abraham Heiftes & Izhar Wallach. First CNN for docking rescoring	Exscientia & DSP launch first Ph I of AI de-novo designed drug	Insilico Medicine announces first-in-class fibrosis drug (TNIK inhibitor INS018_055) to reach PhII

Generations

VANTAI

1981



Xu, The path to the next computational transformation of drug discovery, Medium, 2022

VANTAI

However, we should not forget that **irrational** drug discovery is incredibly successful

haloperidol
1958

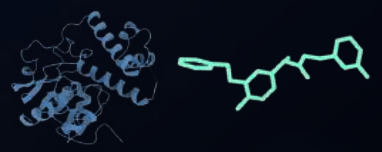
”It was a matter of life or death — a matter of survival. **I don’t believe we really had a clear-cut strategy. We were simply doing whatever we could, and there weren’t many things that we could do then**. We didn’t have much money and there were not many researchers. **We had to make a lot of simple compounds as quickly and possible and screen them using very simple methods”**

Paul Janssen

VANTAI

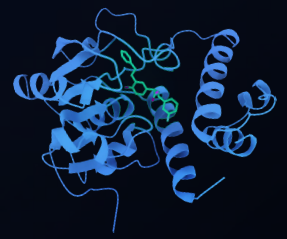
The classical modelling tools focus to solve one task

Structure + Molecule/



Small molecule docking

Molecule

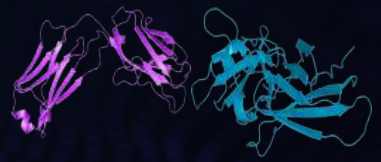


- Optimized to solve one task efficiently, eg. docking: sampling + scoring
- Relies on specific (often rigid) input, that may be prohibitively expensive (eg. X-ray structure) and limited by applicability

Monomers

Protein-Protein Docking

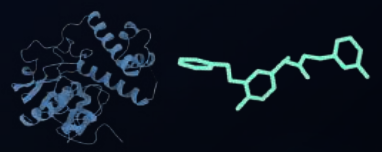
Complex



VANTAI

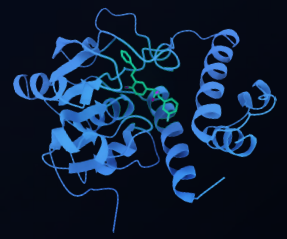
The classical modelling tools focus to solve one task

Structure + Molecule/



Small molecule docking

Molecule

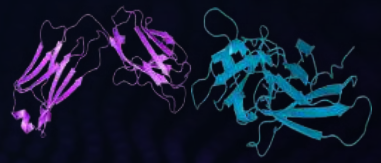


- Optimized to solve one task efficiently, eg. docking: sampling + scoring
- Relies on specific (often rigid) input, that may be prohibitively expensive (eg. X-ray structure) and limited by applicability

Monomers

Protein-Protein Docking

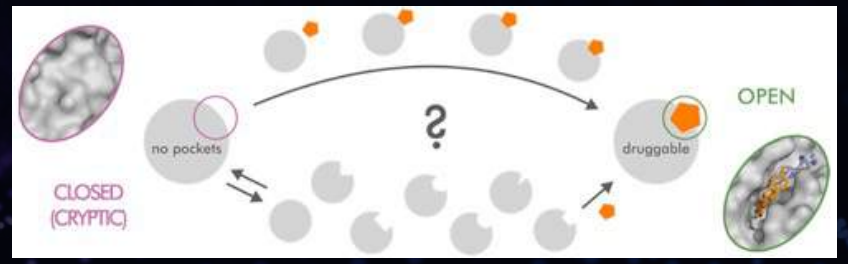
Complex



- Scaling with Moore's law:

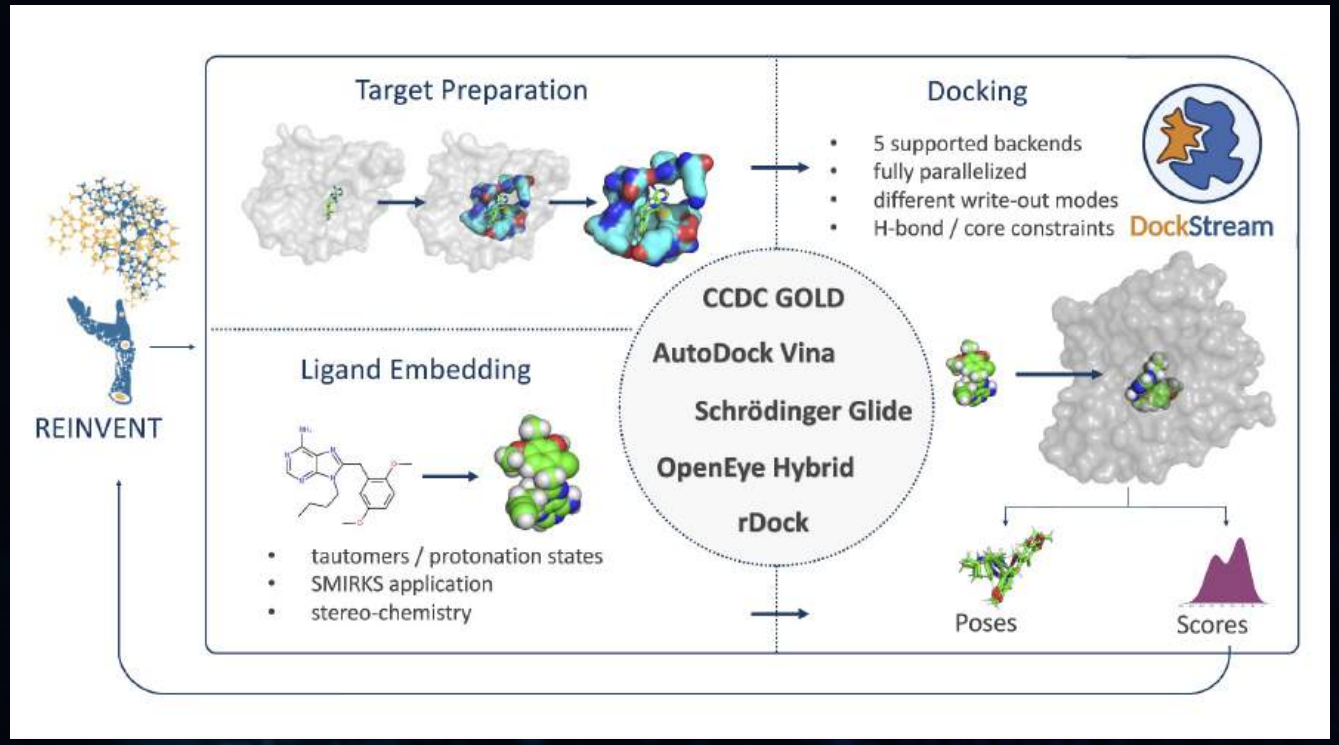
✅ run more iterations, larger libraries / systems

❌ sampling limited by inputs (esp. receptor flex.)



VANTAI

Complex methods limited to sequential combination of tasks



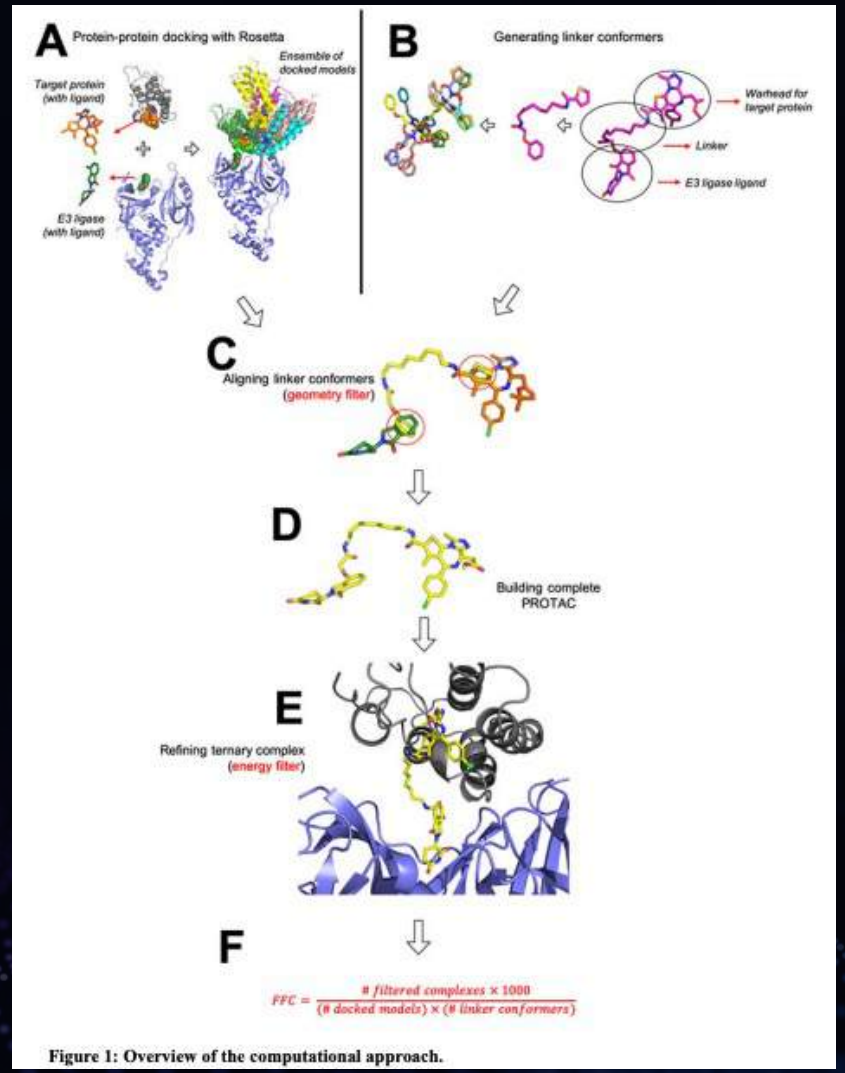


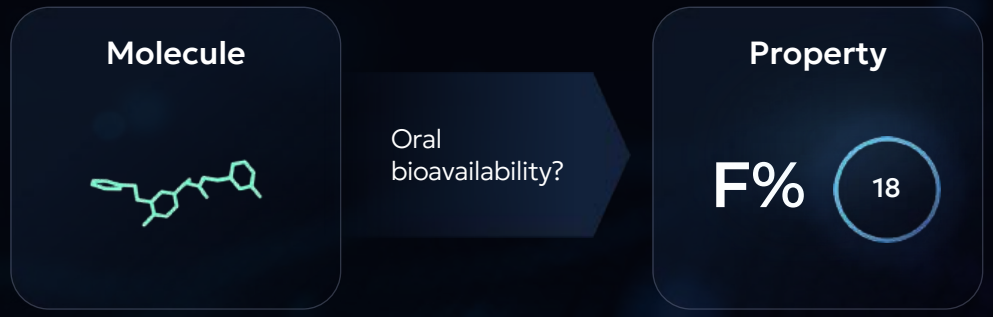
Figure 1: Overview of the computational approach.

https://github.com/MolecularAI/DockStream
https://pubs.acs.org/doi/10.1021/acs.jcim.0c01451

VANTAI

There's broadly two flavors of machine learning that are used

Discriminative

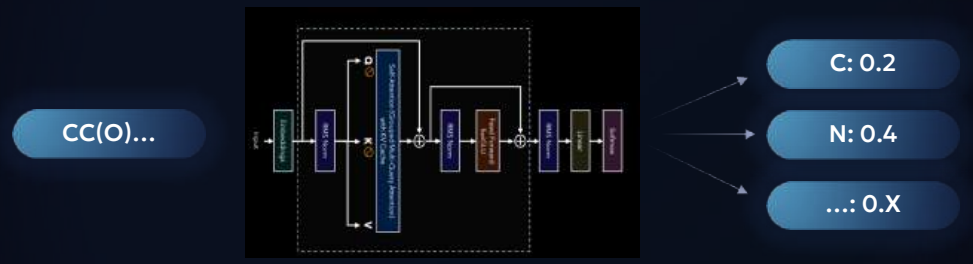


Generative

VANTAI

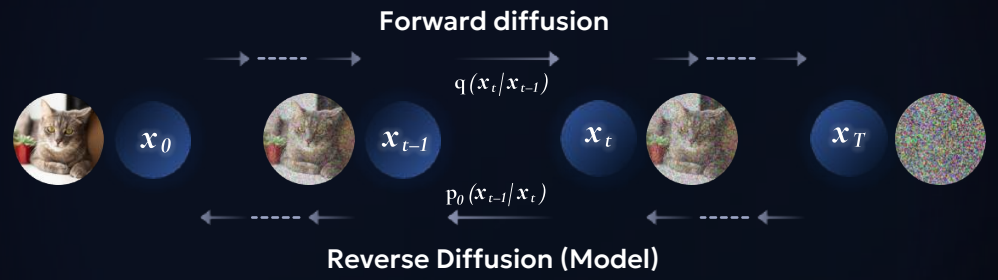
Two methods have been particularly critical in the emergence of generative methods

Language models



- Trained to predict token: either next token (e.g. ChatGPT, “autoregressive”) or some masked tokens in middle of sequence (e.g ESM3)
- Breaks down problem into simpler steps: one token at a time
- Tokens require discrete data
- Became SOTA for language ~= with GPT-1 in 2018

Diffusion models



- Trained to remove simple noise added during training process
- Breaks down problem into simpler steps: only predict ”a bit” of noise – makes problem easier since e.g. high and low noise steps often very different problems
- Usually used for continuous data: images, coordinates, expression levels, …
- Usually a lot more fancy math and many extensions (Flows, Schrodinger Bridges, …)
- Became SOTA for images ~= with ADM in 2021

VANTAI

Diffusion: A breakthrough approach for generative modelling

- Recent breakthroughs leverage Diffusion models
- Inspired by physical diffusion processes (brownian motion)
- Unsupervised training - stepwise noise addition to ground truth
- Model learns going from noised sample to ground truth
- Noise depends on problem at hand – Gaussian noise on pixels, latent noise or more complex (SO(3)-subspace diffusion)

OpenAI DALL·E 3 | Stable Diffusion | Midjourney | Imagen

Generative Adversarial Networks (2015)

Images of flowers

GANs & Roses

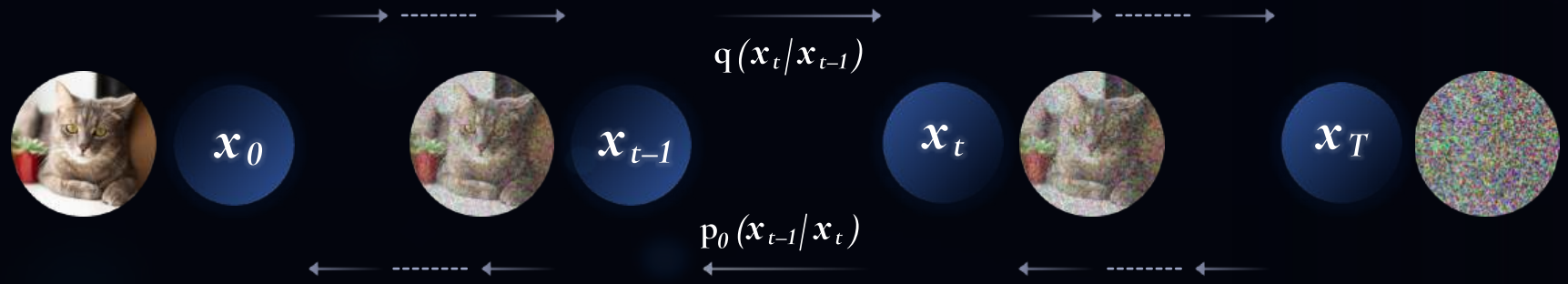
Diffusion Models (2020)

"Show a molecular glue holding together two planets "

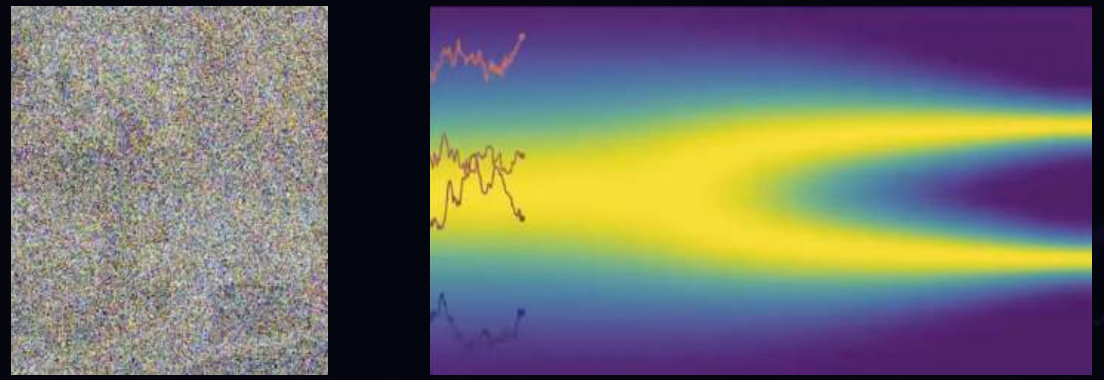
VANTAI

Diffusion Models are trained unsupervised by adding noise to training data and predicting ground truth

Forward diffusion



Reverse Diffusion (Model)

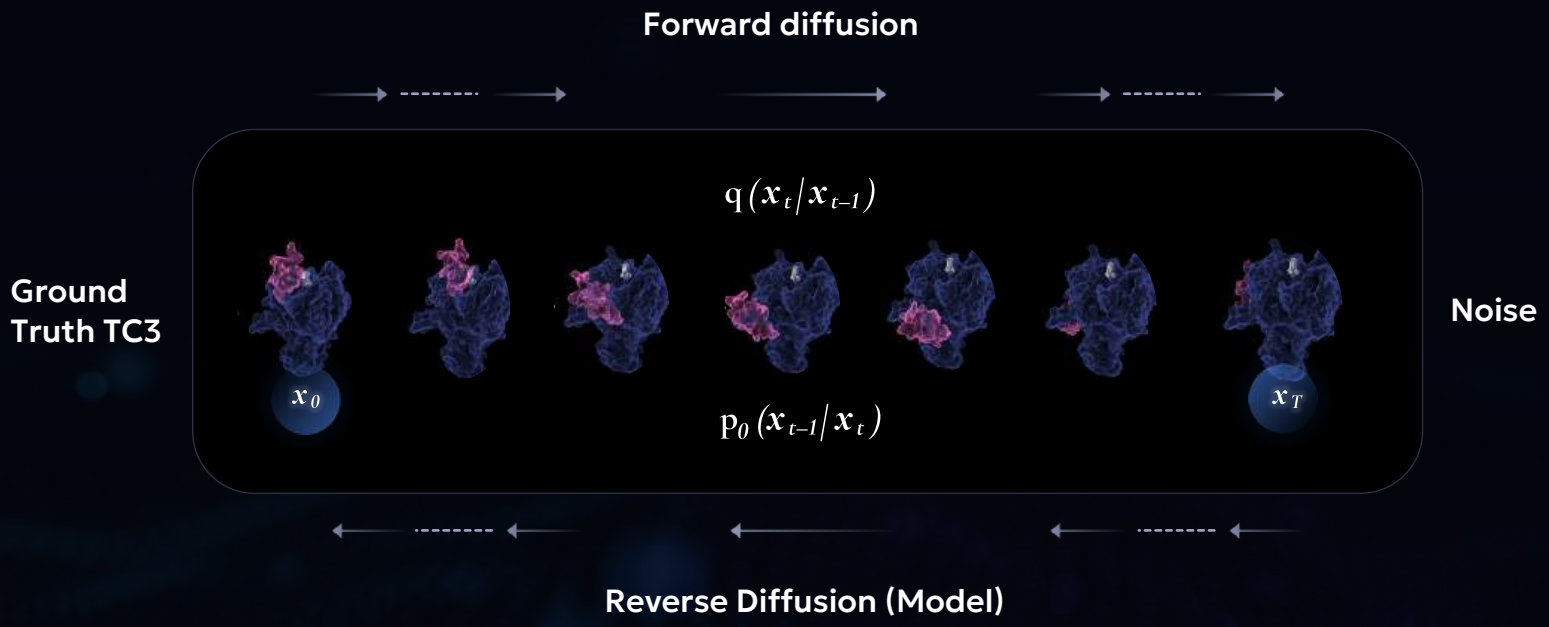


Diffusion probabilistic models - Jascha Sohl-Dickstein, Google Brain Talks
https://www.youtube.com/watch?v=XCUlnHP1TNM

Generative Modeling by Estimating Gradients of the Data Distribution. (NeurIPS 2019) by Yang Song, Stefano Ermon

VANTAI

In molecular modelling, “noise” can be added in highly flexible ways



Noise types

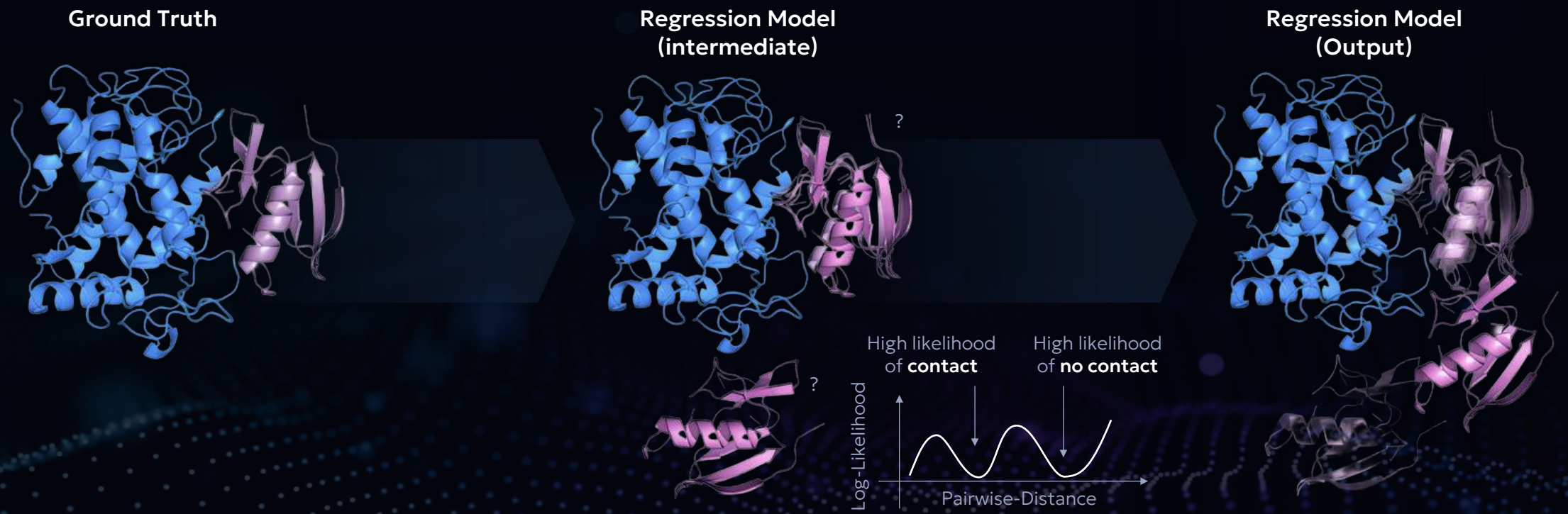
- Example: SE(3) diffusion (rotation, translation): **DiffMaSIF**
- Euclidean diffusion: **ApolloDiff**
- Latent Diffusion: **LatentDiff**

Generative Modeling by Estimating Gradients of the Data Distribution. (NeurIPS 2019) by Yang Song, Stefano Ermon

VANTAI

Why use Diffusion for structure prediction problems?

Apart from the fact the protein structures are not static,
Regression models (e.g. AF2) have a major issue



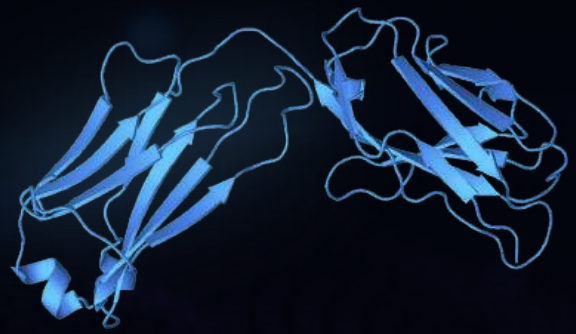
VANTAI

AI is generally used to predict one or multiple of the following given one or multiple of them as input

Sequence

MGRLQLVVLGREDAHFIYENKDVSQ...

Structure



Property

Kinase

Early ML methods mostly followed the classic methods focusing on the same tasks following the same limitations...

VANTAI

This has led to a large ”Zoo” of models in use

Sequence MGRLQLVVLGREDAHFIYENKDVSQ... → Folding → **Structure** → **Structure** → Small molecule generation → **Molecule**

Monomers → Protein-Protein Docking → **Complex** → **Structure + Molecule/** → Small molecule docking → **Molecule**

Structure → Inverse folding → **Sequence** MGRLQLVVLGREDAHFIYENKDVSQ... → **Structure + Molecule/** → QSAR/Property Prediction → **Property** logP 4

Sequence MGRLQLVQ + CC(O)CN → Cofolding → **Structure** → **Molecule** → Synthesis prediction → **Synthesis** A + B + C =

VANTAI

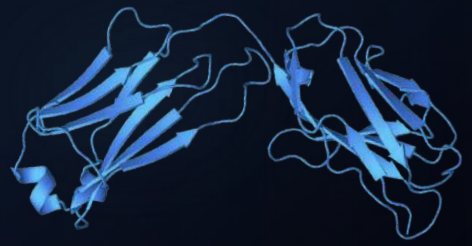
Breakthrough: Protein folding

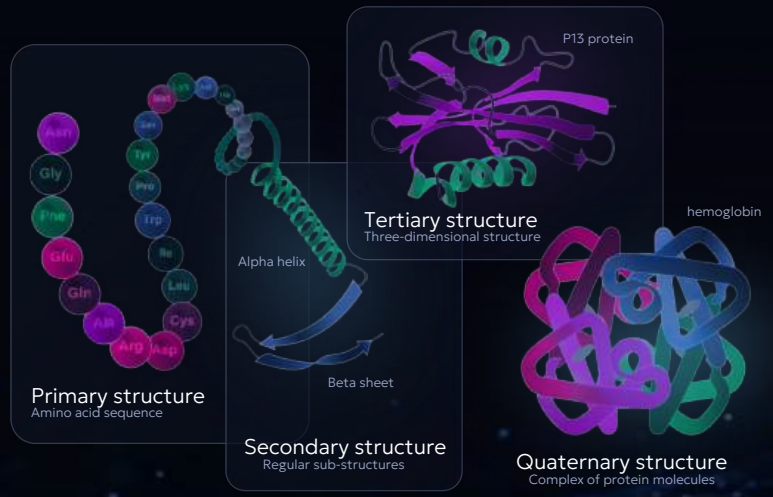
Sequence

MGRLQLVVLGREDAHFIYENKDVSQ...

Folding

Structure





A protein's structure is uniquely determined by its sequence

"Anfinsen's Dogma" – 1972 Nobel Prize

VANTAI

Traditional route:

X-ray crystallography

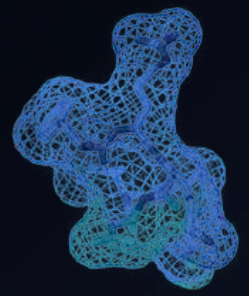
Crystal

→ x-rays

Diffraction pattern

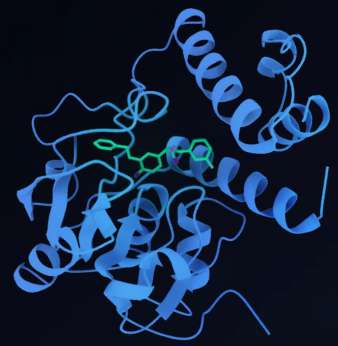
→ phases

Electron density map



→ fitting

Atomic model

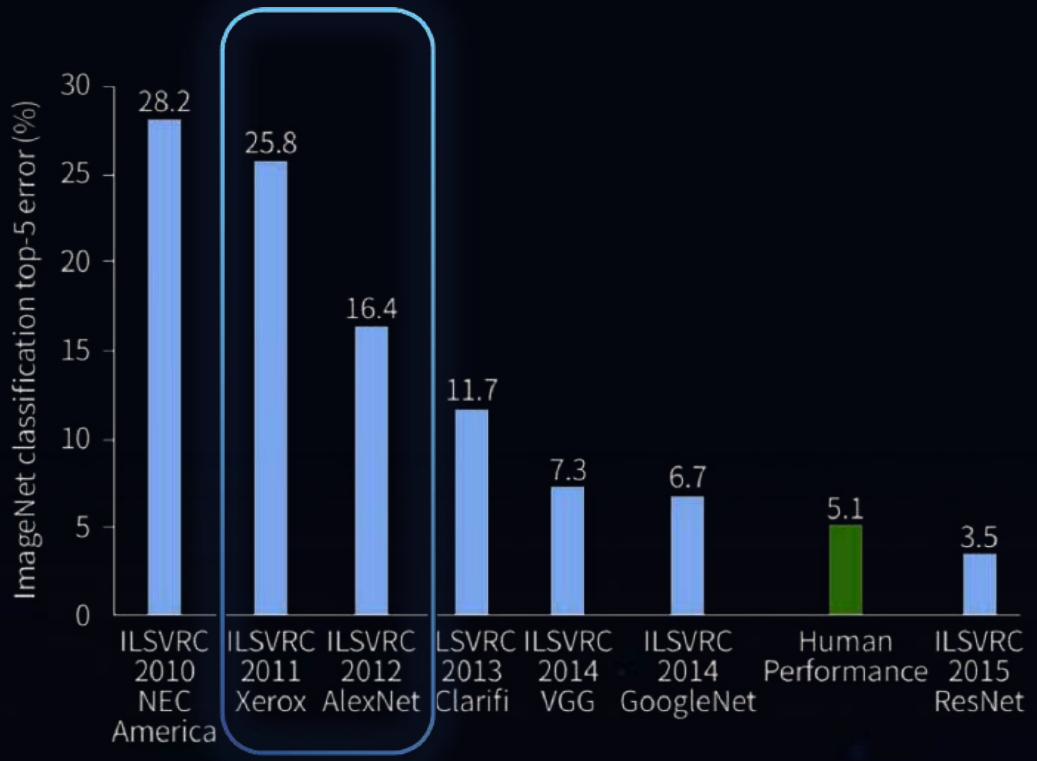


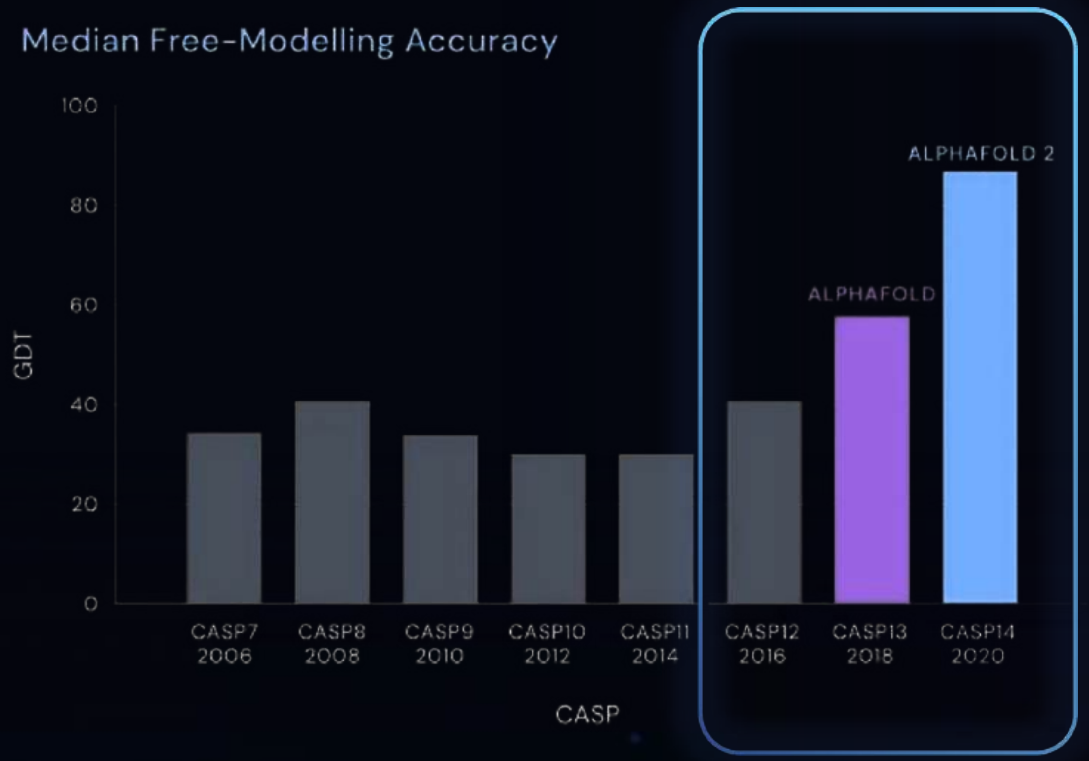
Up to **1M USD**

Can take **years**

VANTAI

AF2 was an “AlexNet” like moment



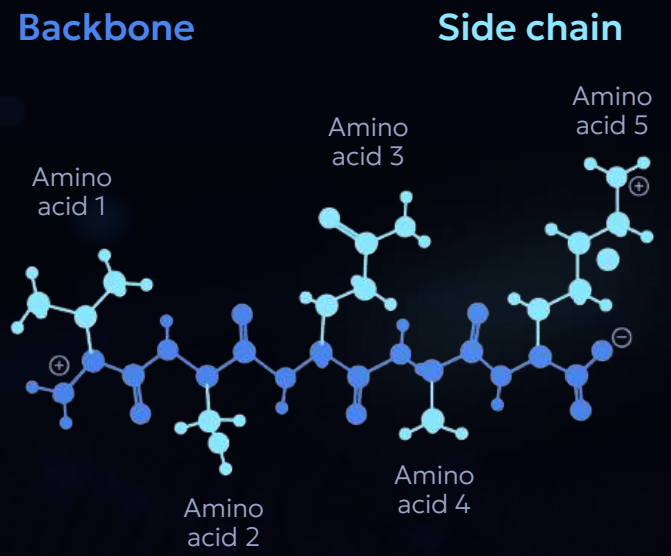


Amy Lu, BIOE 145/245: Introduction to Machine Learning for Computational Biology,Spring 2024

VANTAI

Proteins are polymers comprised of amino acids

›UniRef9O_P02057 Hemoglobin subunit beta-1/2 n=6 Tax=Euarchontoglires TaxID=314146 RepID=HBELRABIT
MVHLSSEEKSAVTALWGKVNVEEVGGEALGRLLVVYPWT
QRFFESFGDLSSANAVMNWKVKAHGKKVLAAFSEGLSHL
DNLKGTFAKLSELHCDKLHVDPENERLLGNVLVIVLSHHFG
KEFTPQVQAAYQKVVAGVANALAHKYH

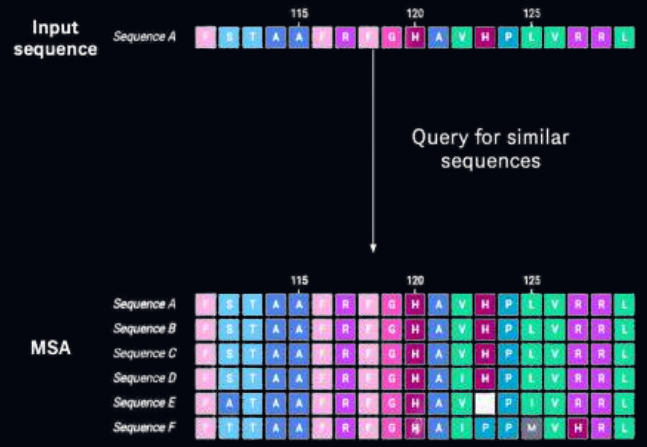


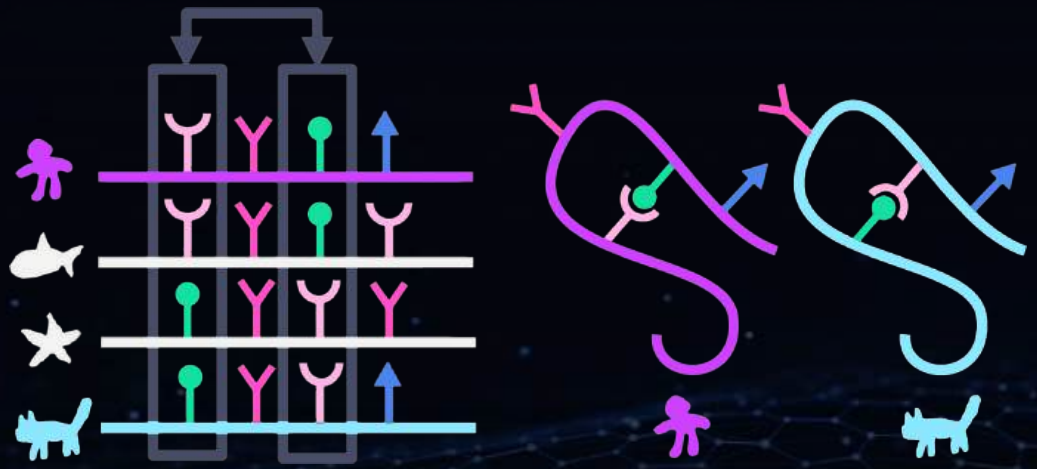
- We usually split structures in backbone (mainchain) & sidechain
- Backbone structure often assumed +/- independent of specific sidechains
- E.g. for protein design: often design backbones and then predict different residues that could take this shape

Amy Lu, BIOE 145/245: Introduction to Machine Learning for Computational Biology,Spring 2024

VANTAI

However, technically AlphaFold2 doesn't fold protein sequences, it predicts structures based on Multiple Sequence alignment





- MSA gives us co-evolution
- Residue pairs that co-evolve (e.g charge switch on one always observed with opposing charge switch on other) are likely in contact
- Often ignored: it's also a form of **RAG (retrieval augmented generation)** – giving large number of sequences that likely fold into same structure makes sequence-based fold retrieval more robust

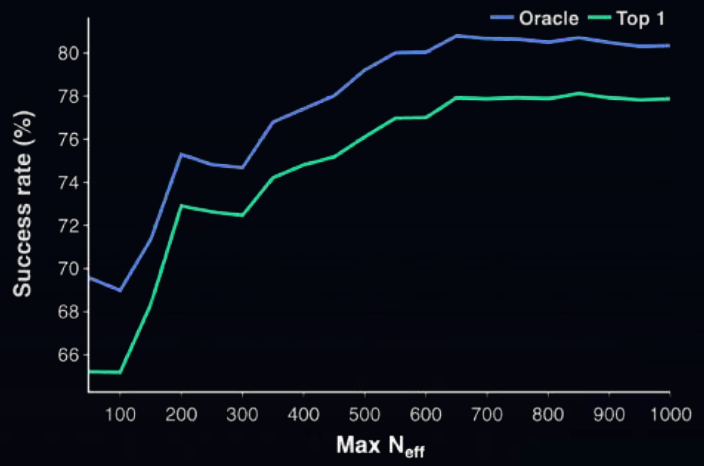
Amy Lu, BIOE 145/245: Introduction to Machine Learning for Computational Biology,Spring 2024

VANTAI

In fact, co-evolutionary information is absolutely critical

Source: Zeming Lin et al., 2021

AF2 PINDER performance by MSA depth



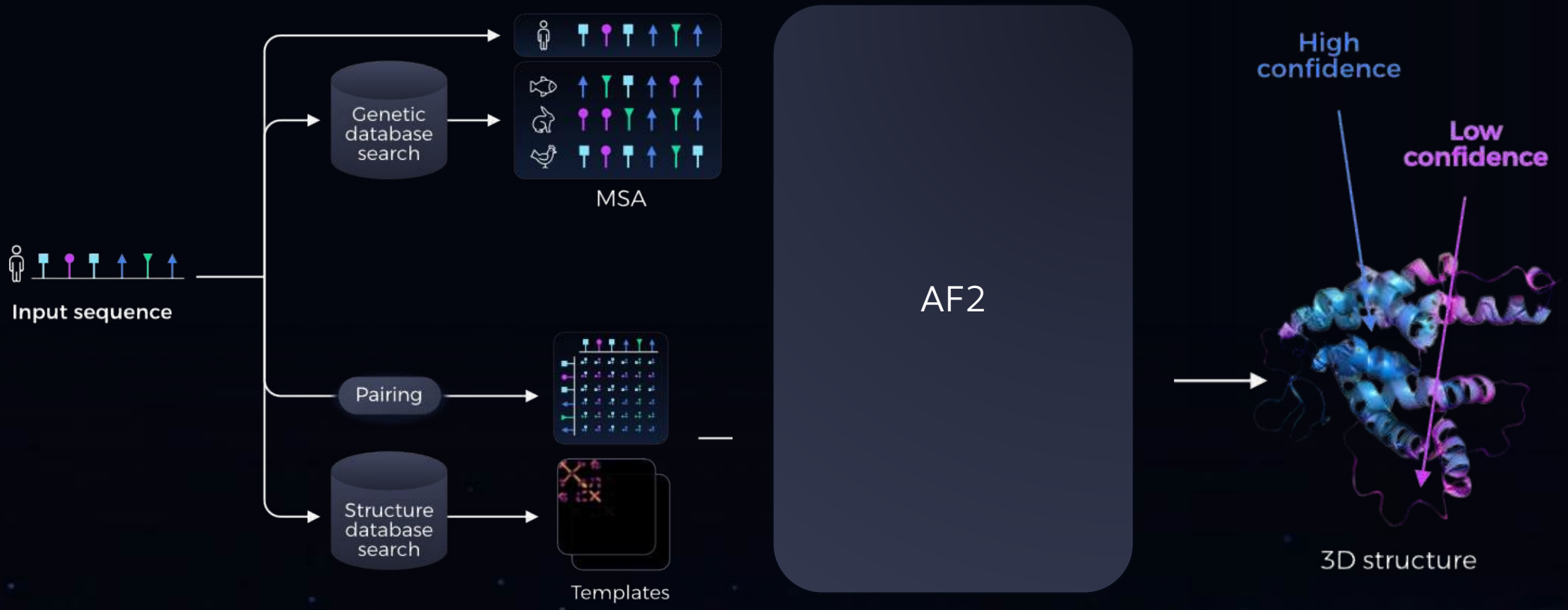
Source: Kovtun et al. PINDER, 2024 (VantAI)

Notes:

- For both monomers and protein-protein interfaces, MSA are absolutely critical information
- While current language models such as ESM3 have been shown to learn the pairwise residue covariance implicitly, despite 100B+ they still underperform a sequence alignment

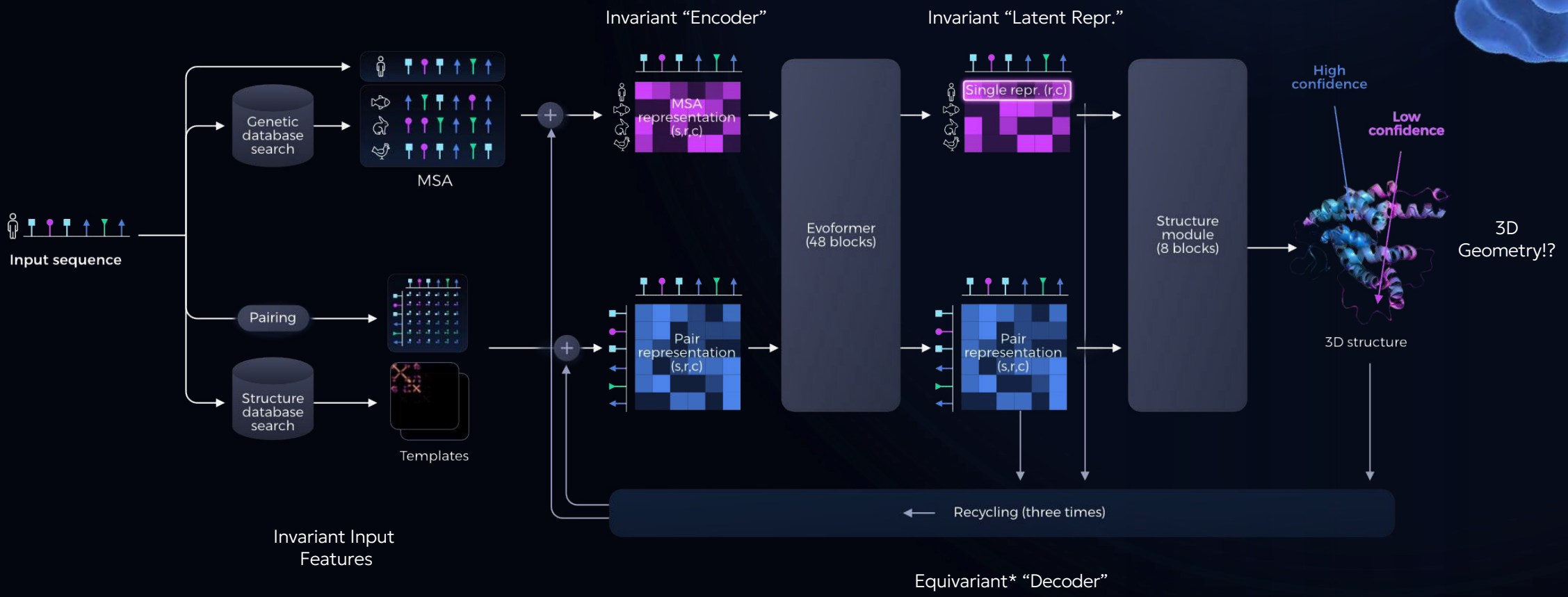
VANTAI

AF2 is a model that uses sequence and co-evolution to predict a protein's structure



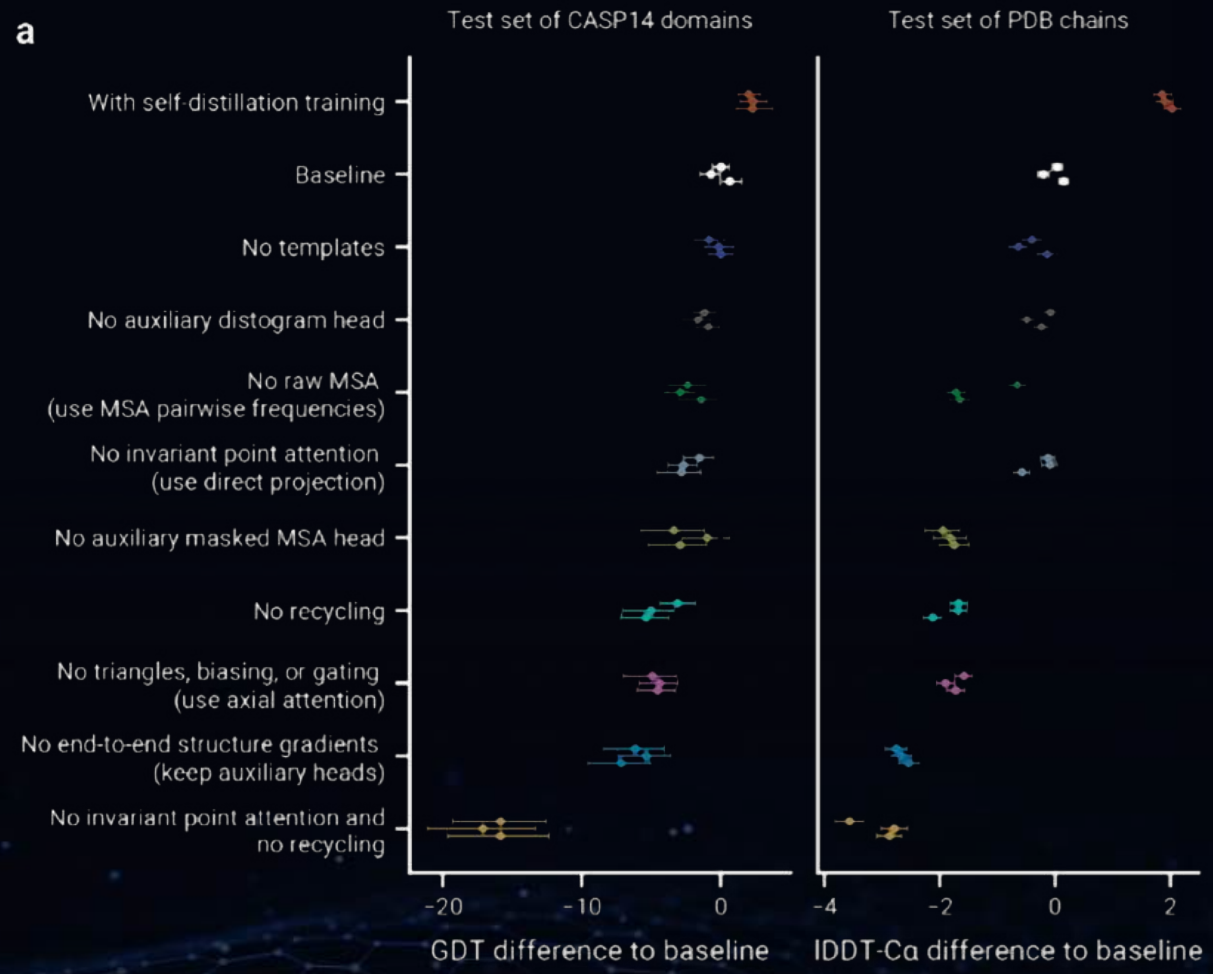
VANTAI

It uses a transformer inspired architecture to first process sequence & MSA features and then reason over coordinates



VANTAI

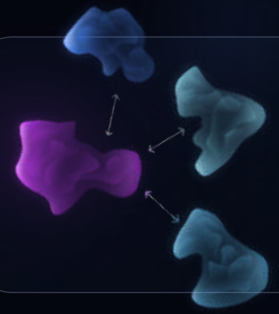
Critically – AF2 is an “engineering marvel” with many, many performance critical innovations



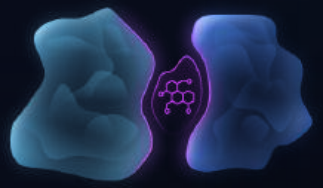
VANTAI

De-novo molecular generators for variety of use cases

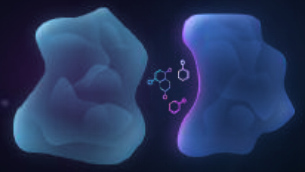
Example use-cases
for Proximity Modulators:



De-novo generative binder or glue design



Generative Linker Design



Generative Lead opt

Generative Algorithms

Diffusion/Flow Matching
Probabilistic Models

LLMs +
RL/RLHF

Genetic algorithms
or search-based
generative methods

Scoring functions

Shape-based

Free Energy
& interaction-based

ADME/PK
Property classifiers

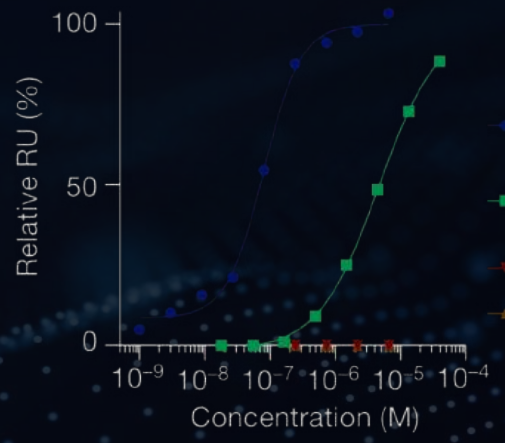
VANTAI

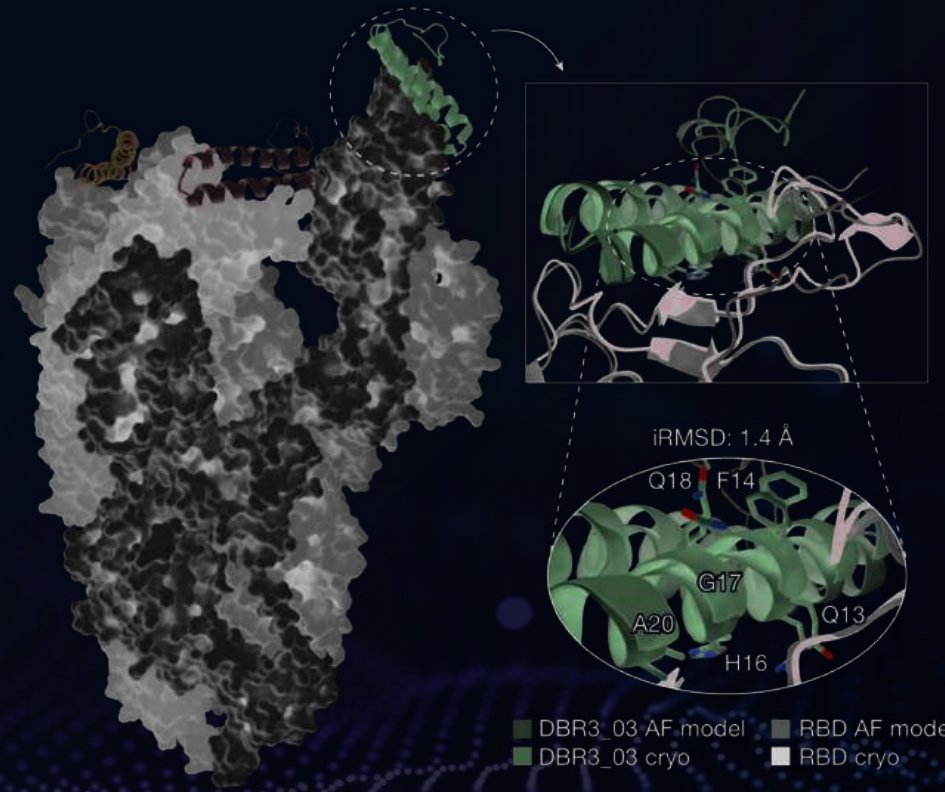
MaSIF & dMaSIF have been highly successful for protein-design applications

De novo design of protein interactions with learned surface fingerprints

https://doi.org/10.1038/s41586-023-05993-x
Received: 16 June 2022
Accepted: 21 March 2023
Published online: 26 April 2023
Open access
Check for updates

Pablo Gainza[1,2,10,11], Sarah Wehrle[1,2,11], Alexandra Van Hall-Beauvais[1,2,11], Anthony Marchand[1,2,11], Andreas Scheck[1,2,11], Zander Harteveld[1,2], Stephen Buckley[1,2], Dongchun Ni[3,4], Shuguang Tan[6], Freyr Sverrisson[1,2], Casper Goverde[1,2], Priscilla Turelli[6], Charlène Raclot[6], Alexandra Teslenko[7], Martin Pacesa[1,2], Stéphane Rosset[1,2], Sandrine Georgeon[1,2], Jane Marsden[1,2], Aaron Petruzzella[8], Kefang Liu[5], Zepeng Xu[5], Yan Chai[5], Pu Han[5], George F. Gao[5], Elisa Oricchio[8], Beat Fierz[7], Didier Trono[6], Henning Stahlberg[3,4], Michael Bronstein[9,10] & Bruno E. Correia[1,2]





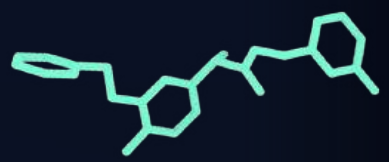
VANTAI

A variety of generative algorithms have been used, usually split into 2D or 3D representations

Generative Algorithms

- Diffusion/Flow Matching Probabilistic Models
- LLMs + RL/RLHF
- Genetic algorithms or search-based generative methods

3D

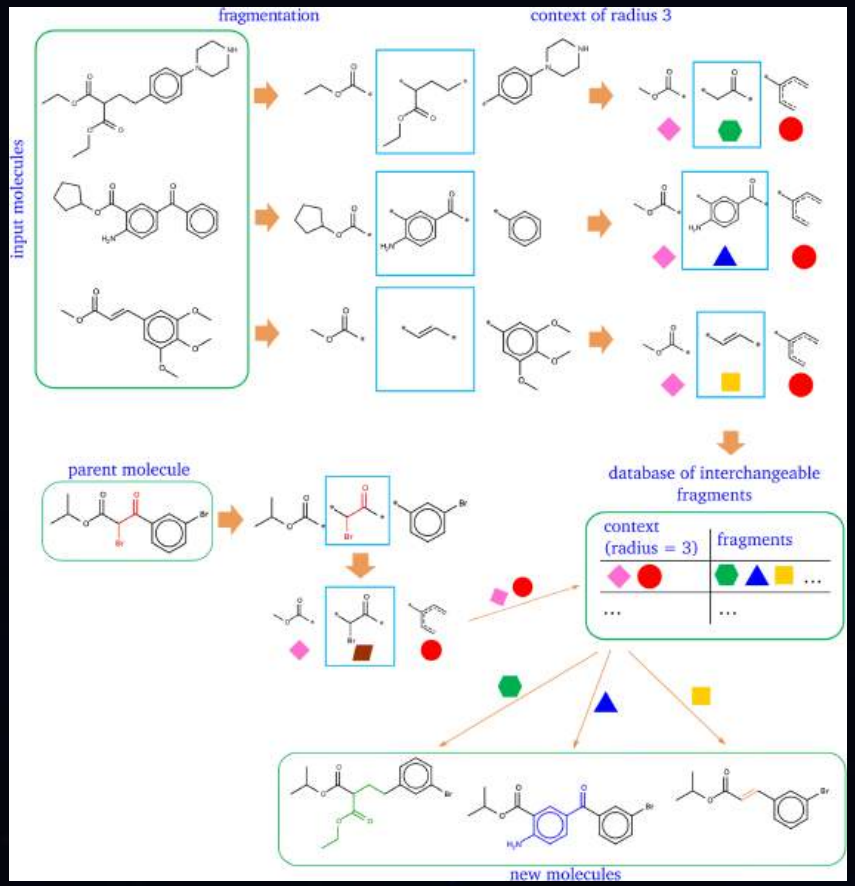


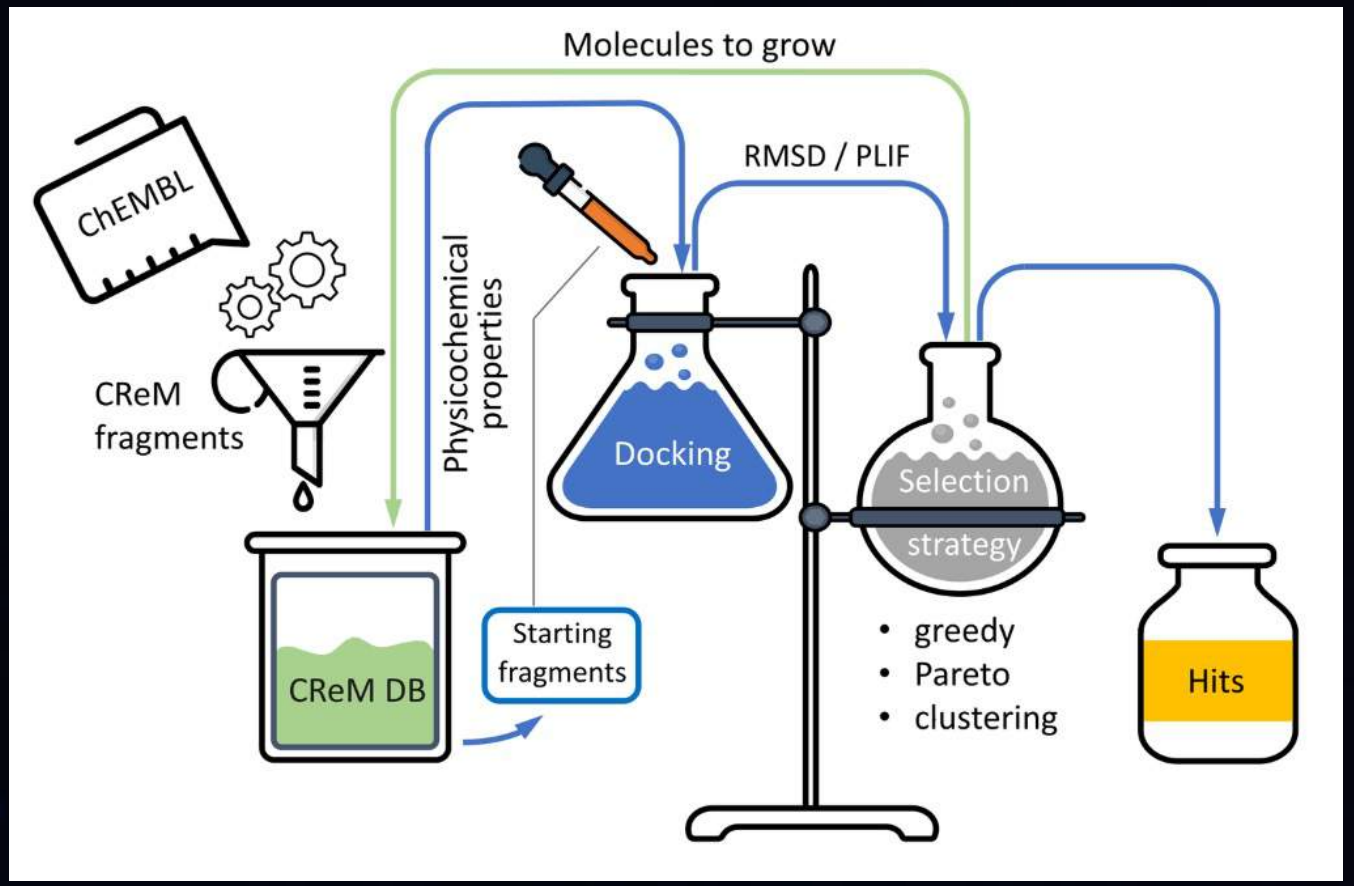
2D

CC(O)CN

VANTAI

Classical approaches: rule-based (iterative) enumerators



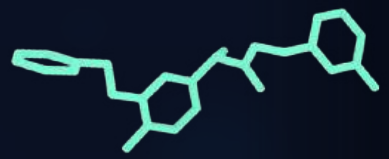


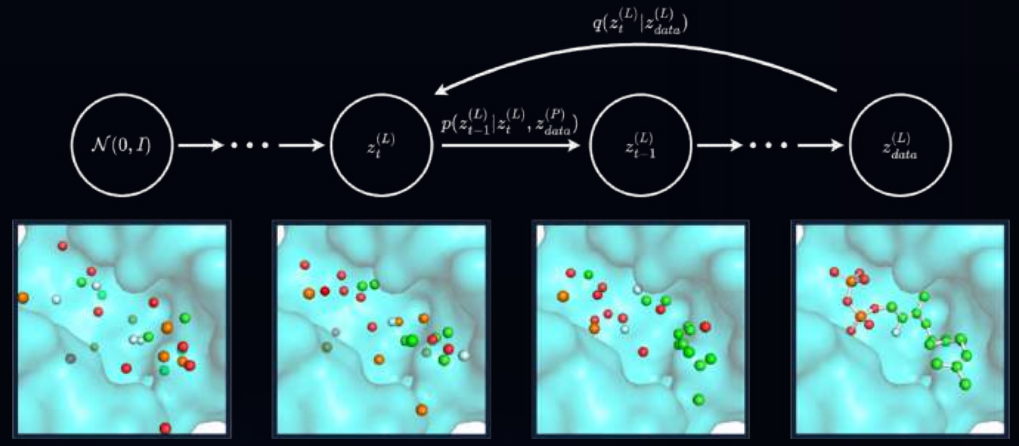
https://github.com/ci-lab-cz/crem-dock

VANTAI

A variety of generative algorithms have been used, usually split into 2D or 3D representations

3D





Structure-based Drug Design with Equivariant Diffusion Models. Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom Blundell, Pietro Lio, Max Welling, Michael Bronstein, Bruno Correia. ArXiV, 2022

2D

CC(O)CN

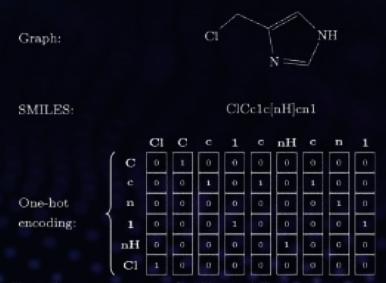


Fig. 3 Three representations of 4-(chloromethyl)-1H-imidazole. Depiction of a one-hot representation derived from the SMILES of a molecule. Here a reduced vocabulary is shown, while in practice a much larger vocabulary that covers all tokens present in the training data is used.

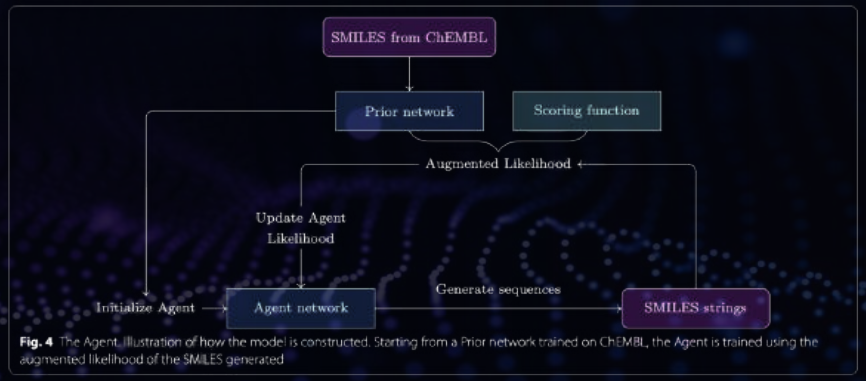
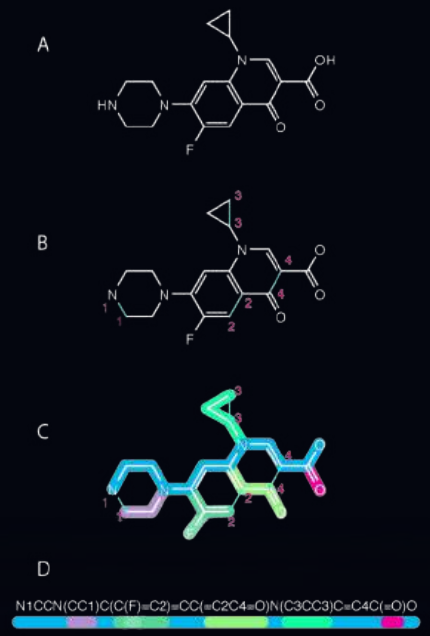


Fig. 4 The Agent. Illustration of how the model is constructed. Starting from a Prior network trained on ChEMBL, the Agent is trained using the augmented likelihood of the SMILES generated.

Molecular de-novo design through deep reinforcement learning
Marcus Olivecrona, Thomas Blaschke, Ola Engkvist and Hongming Chen. J Cheminform (2017)

VANTAI

SMILES is a string representation of a molecular graph which allows standard LLM tokenization approaches



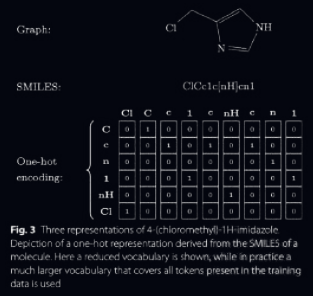


Fig. 3 Three representations of 4-(chloromethyl)-1H-imidazole. Depiction of a one-hot representation derived from the SMILES of a molecule. Here a reduced vocabulary is shown, while in practice a much larger vocabulary that covers all tokens present in the training data is used.

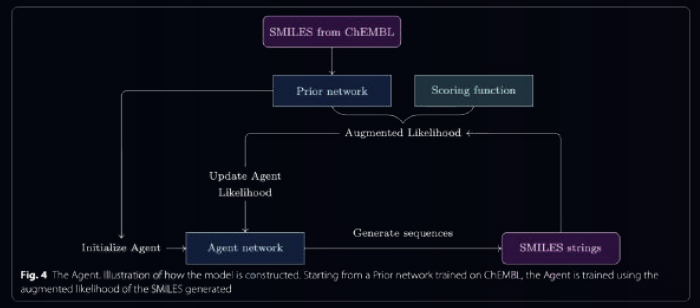
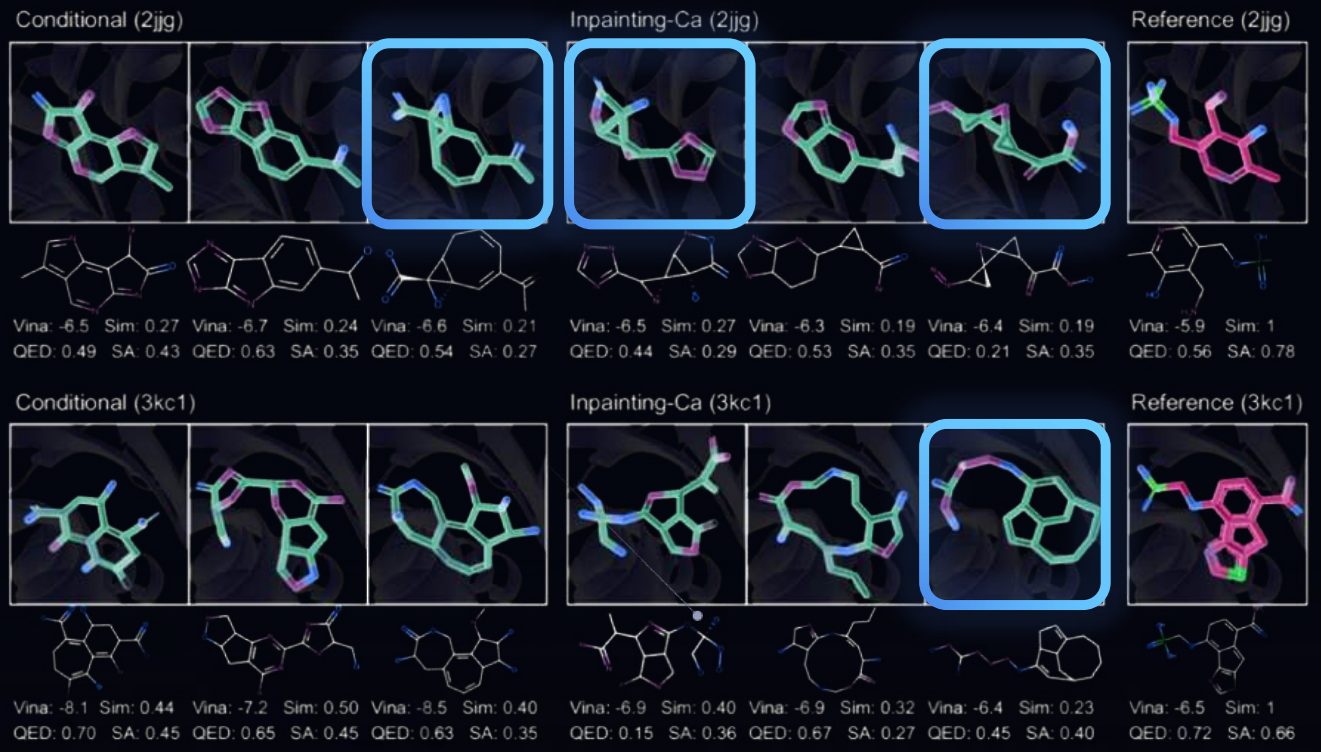


Fig. 4 The Agent. Illustration of how the model is constructed. Starting from a Prior network trained on ChEMBL, the Agent is trained using the augmented likelihood of the SMILES generated.

Molecular de-novo design through deep reinforcement learning
Marcus Olivecrona, Thomas Blaschke, Ola Engkvist and Hongming Chen. J Cheminform (2017)

- SMILES (Simplified Molecular Input Line Entry System) is a depth-first linearized representation of a molecular graph (i.e. converts [cyclic] graphs into a linear sequence of characters expressed in ASCII
- ASCII symbols can then be encoded as in classical LLMs
- Many early RNN-based systems such as REINVENT (AstraZeneca, 2017) which tie SMILES-based LLMs to a learnable reward via RL (REINFORCE) are still competitive today

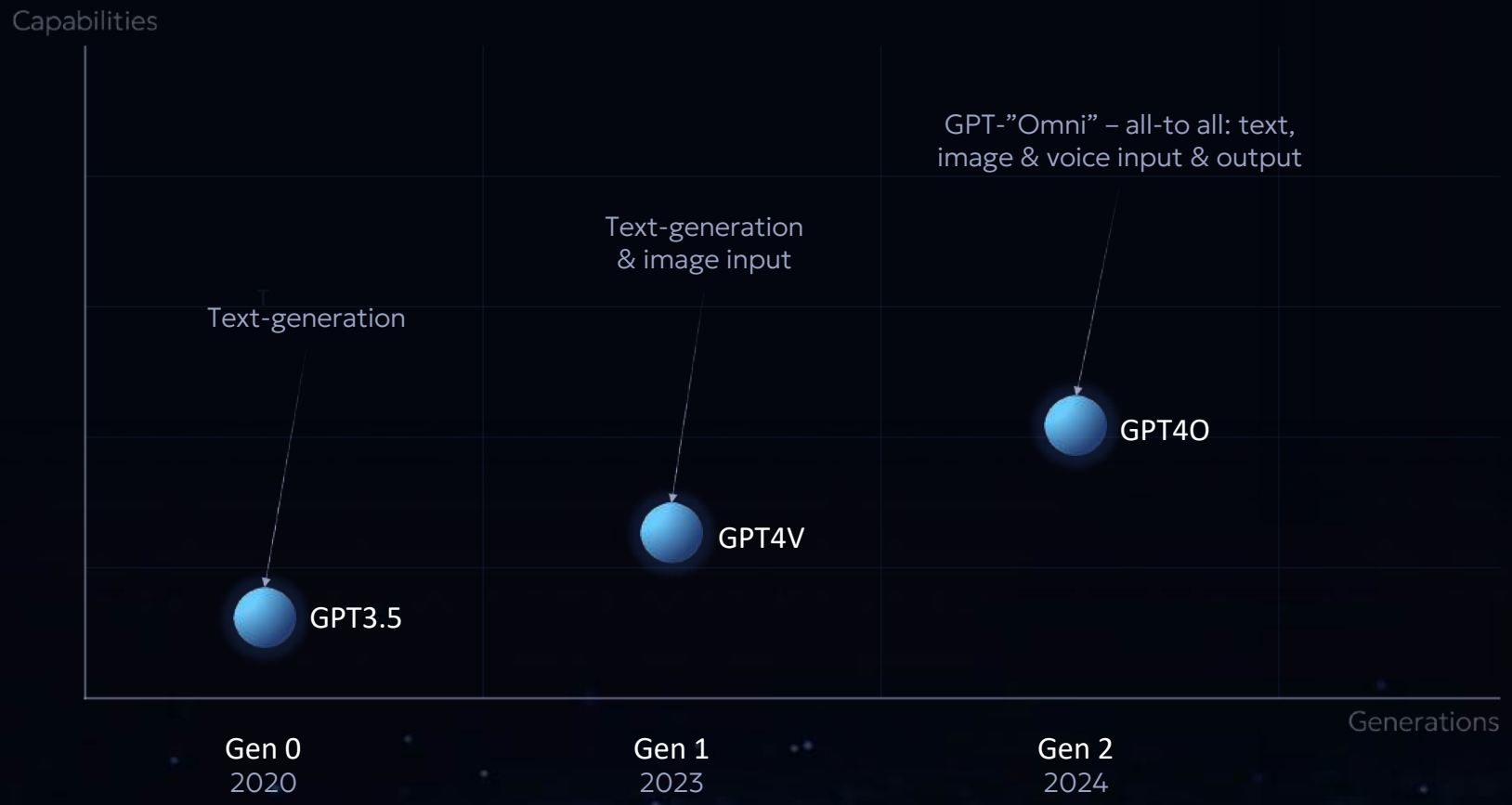
3D-based models still struggle to produce valid molecules



- Strained bond angles
- Highly strained or overextended rings
- Disconnected aromatic bonds
- Large bent overall molecular shapes
- Many methods are still in the “6-Finger phase” of generative models
- Whether same recipe from image world (larger models, more data: scale) will suffice or if explicit inductive biases will be required remains to be seen

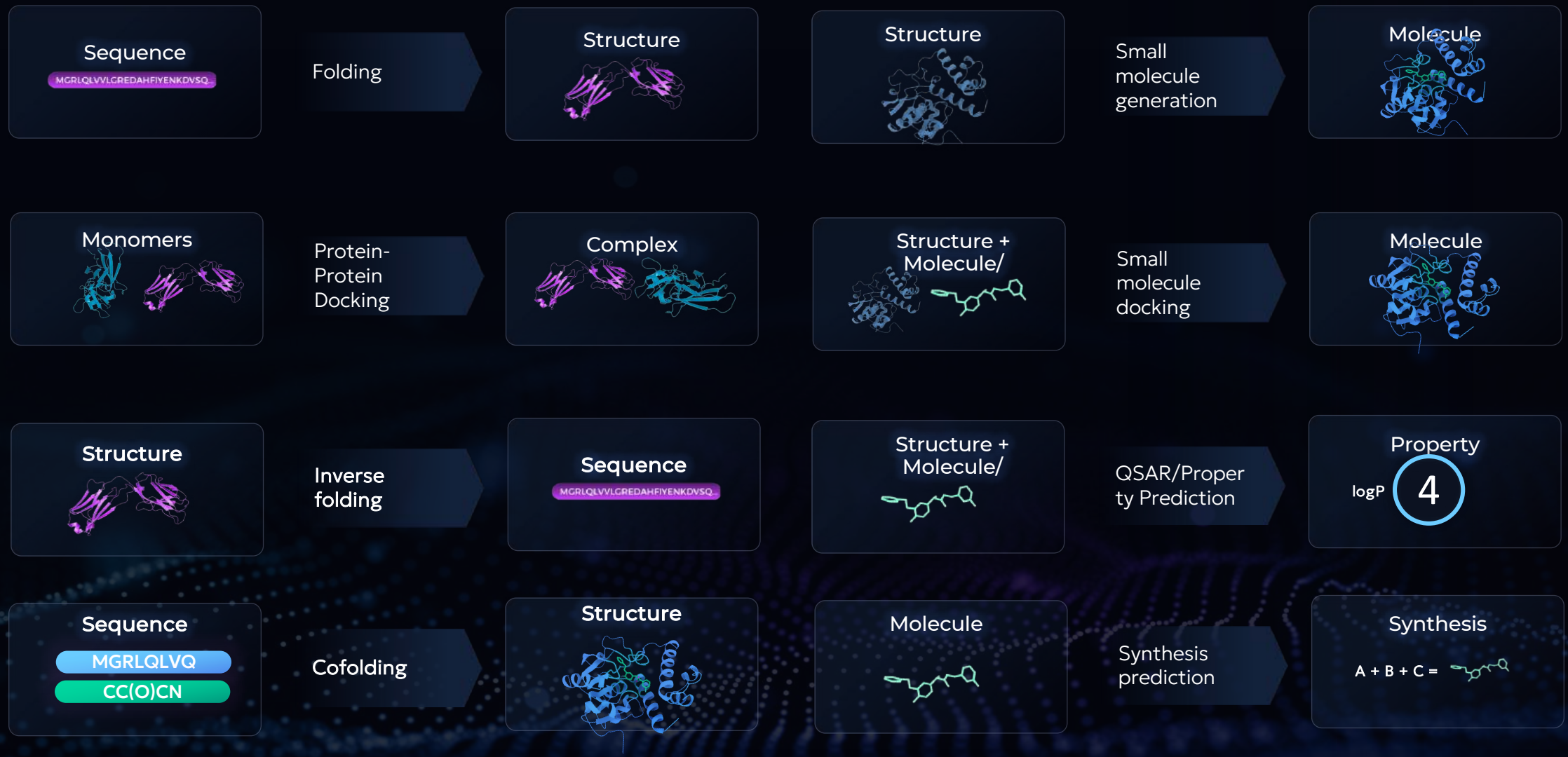
VANTAI

The big breakthroughs in AI have been driven through ever more general models



VANTAI

Current Bio AI models, in contrast, still consist of a large number of specialist models



VANTAI

However, this is wasteful and limiting

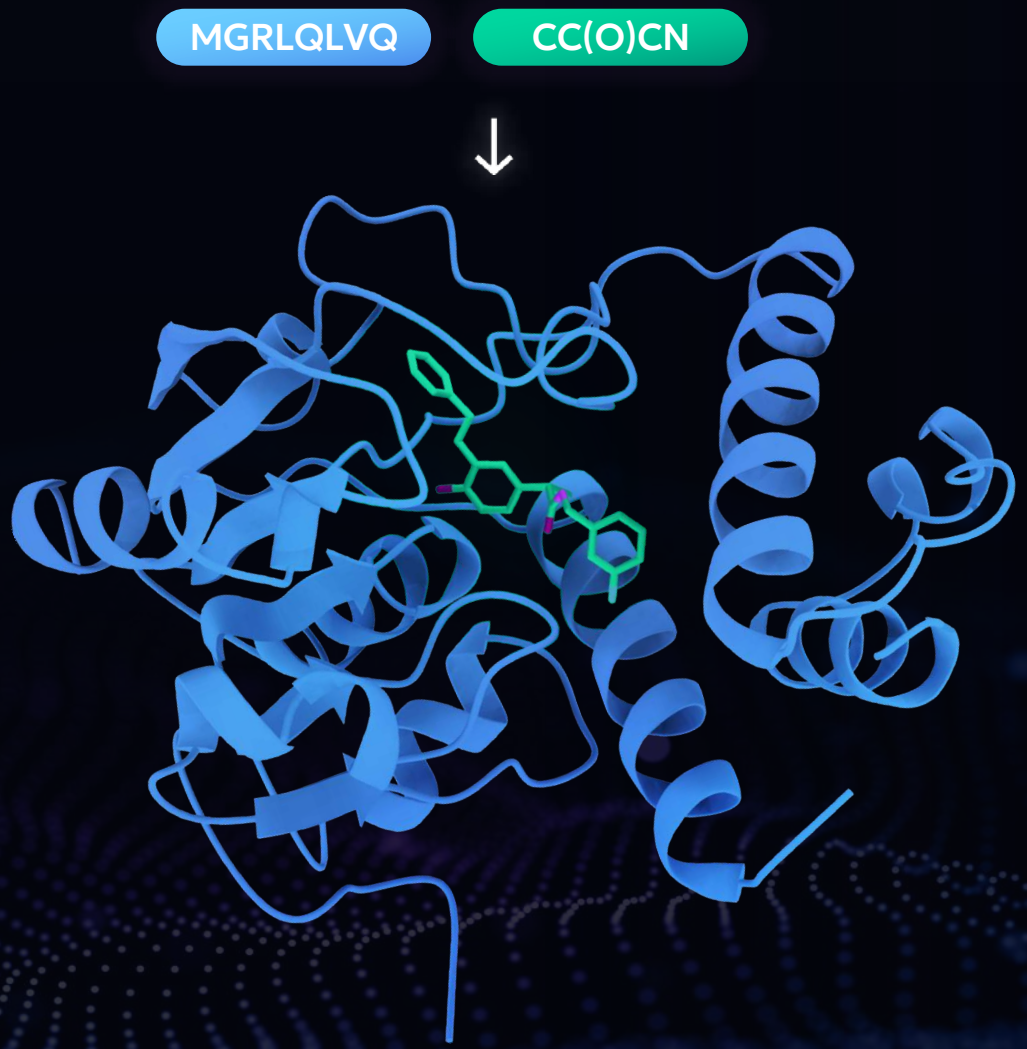
- Models re-learn same things anew (”grammar” of natural sequence, physics of interactions, generative process, …)
- Different and small datasets used to re-learn across tasks
- In real life, we often want to ”feed” models with whatever information we have
- Currently, this requires models to be chained or “hacked” to include information they are not trained to use (e.g. MSA sampling in AF2), creating and compounding errors

With AF3, all atoms of life (DNA, Protein, Molecules) could be predicted for the first time

VANTAI

Co-Folding:

Predict atomistic structure given sequence of molecular tokens



VANTAI

The challenge Neo set out to solve:

Decade+ process of finding an effective molecule

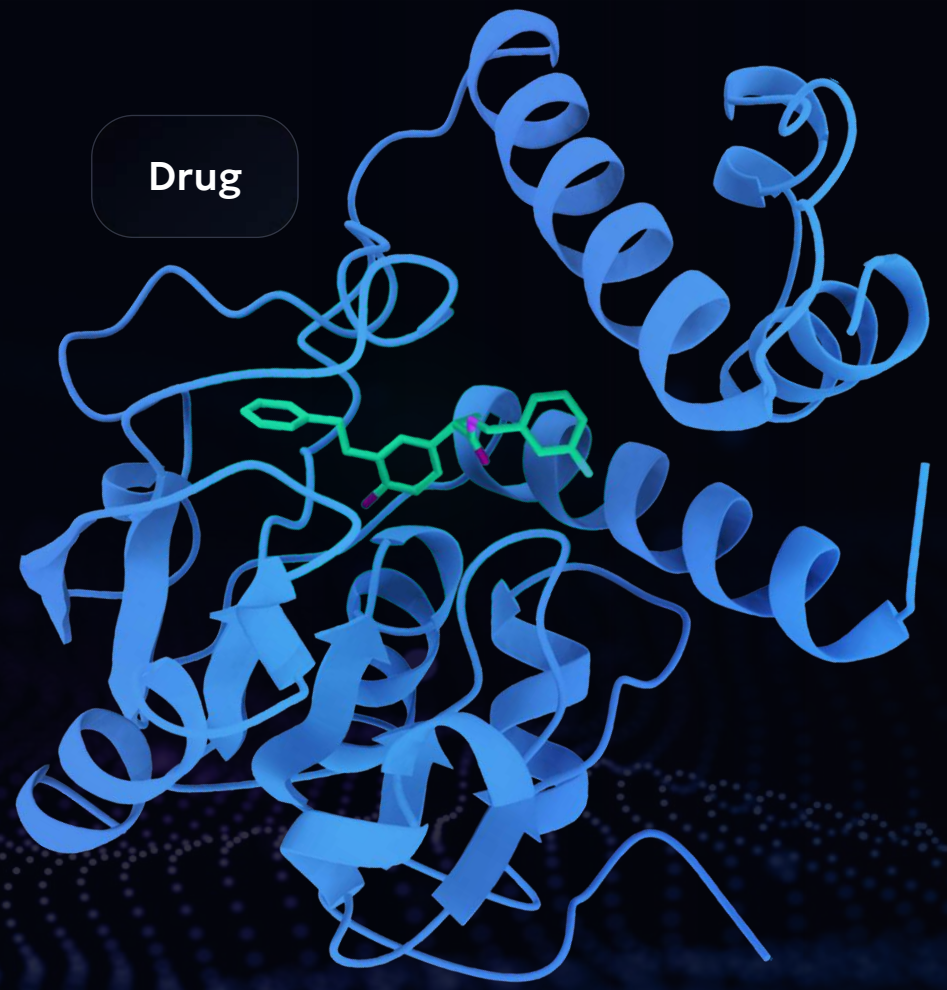
Sequence

MGRLQLVQ

→

13+ years

Drug



VANTAI

Challenge: Design & decode medicines

atom-by-atom

MGRLQLVQ →

1. Design

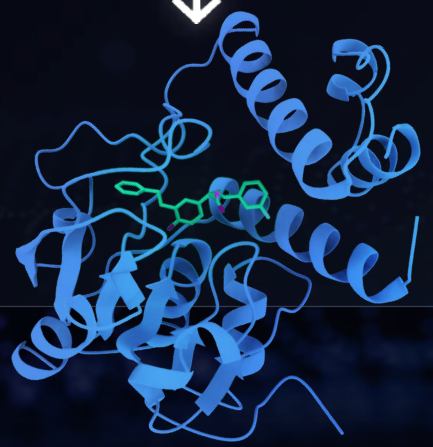
(Molecular generation)

CC(O)CN

2. Decode

(Folding)

MGRLQLVQ CC(O)CN

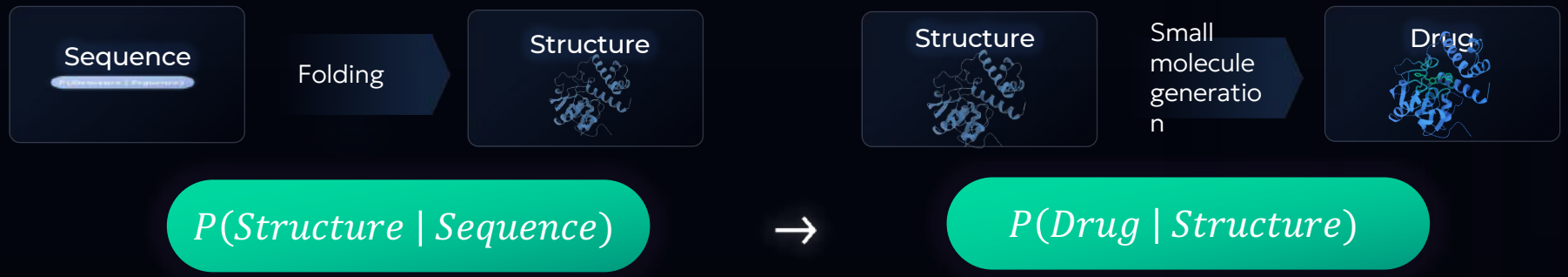


VANTAI

And the next frontier was clear

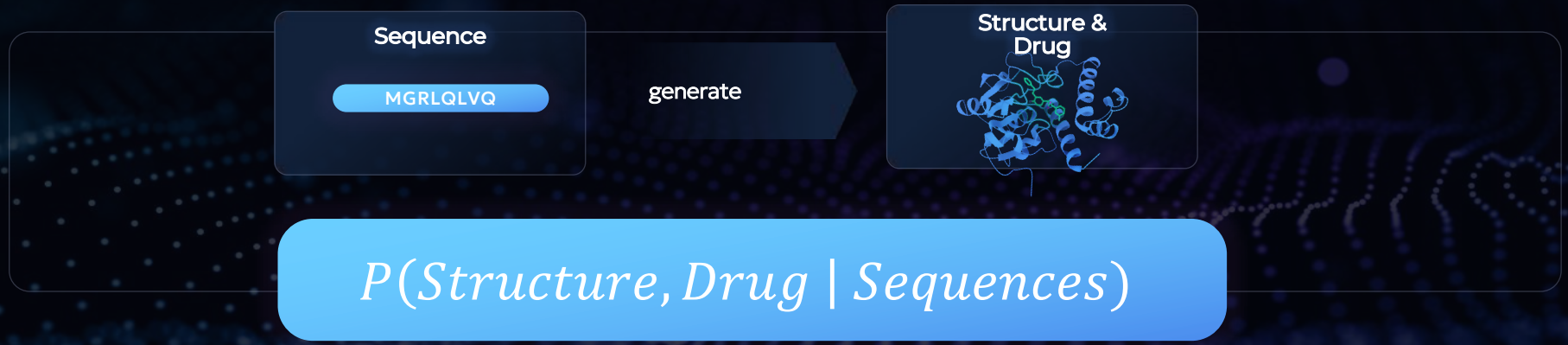
VANTAI

Existing methods



However, in ProMods $Structure = f(Drug, Sequences)$

Needed

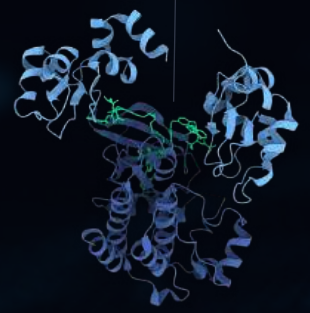


VANTAI

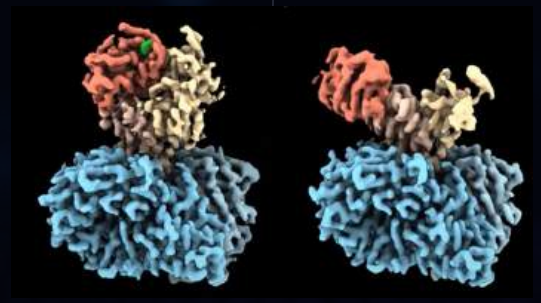
ProMods: molecule defines structure

often **doesn't form** stably **in absence of drug**

Molecule-induced interface

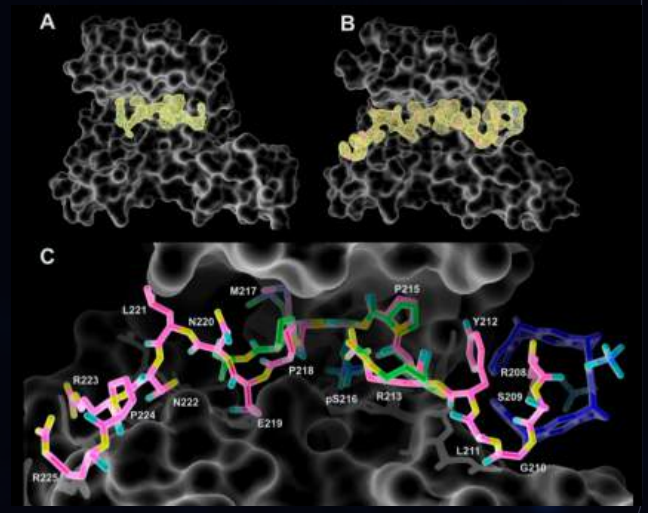


Molecule-induced conformational shift



Pomalidomide-induced conformational shift of CRBN. Gabe Lander, DFCI TPD Series, 2023

Molecule-induced disorder to order conversion



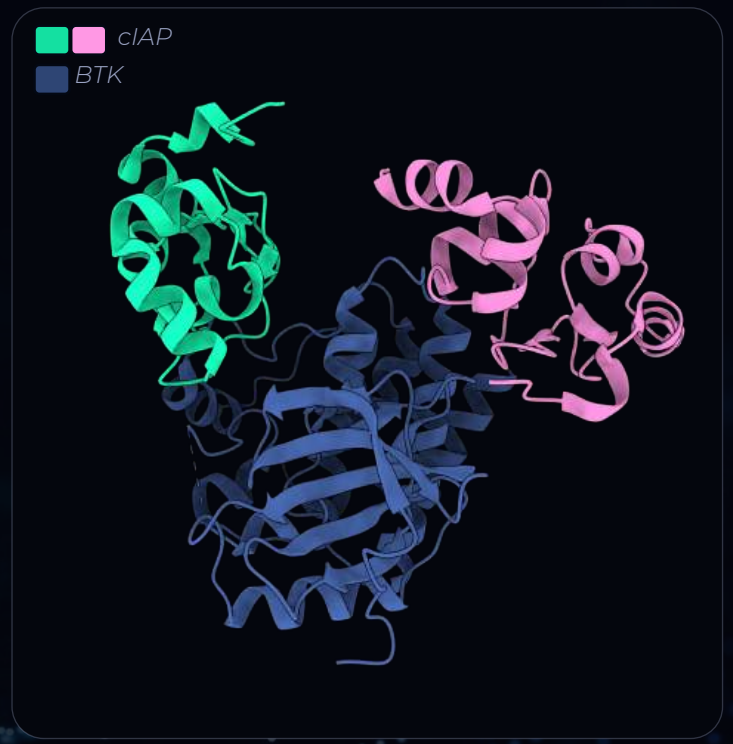
The Molecular Tweezer CLR01 Stabilizes a Disordered Protein-Protein Interface. Bier et al. JACS 2017

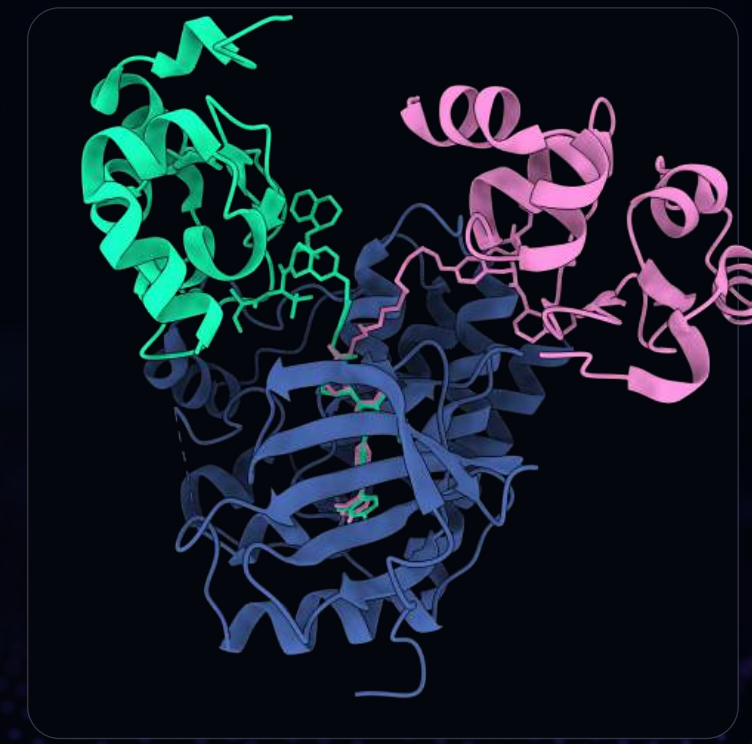
Protein-Protein & Protein-Ligand docking invalidly assume interface and monomers exist stably without molecule – simultaneous co-folding & de-novo design required

$$Structure = f(Drug, Sequences)$$

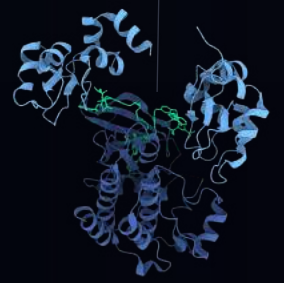
VANTAI

Which of these two cIAP-BTK Protein-interfaces is the right one?





Molecule-induced interface

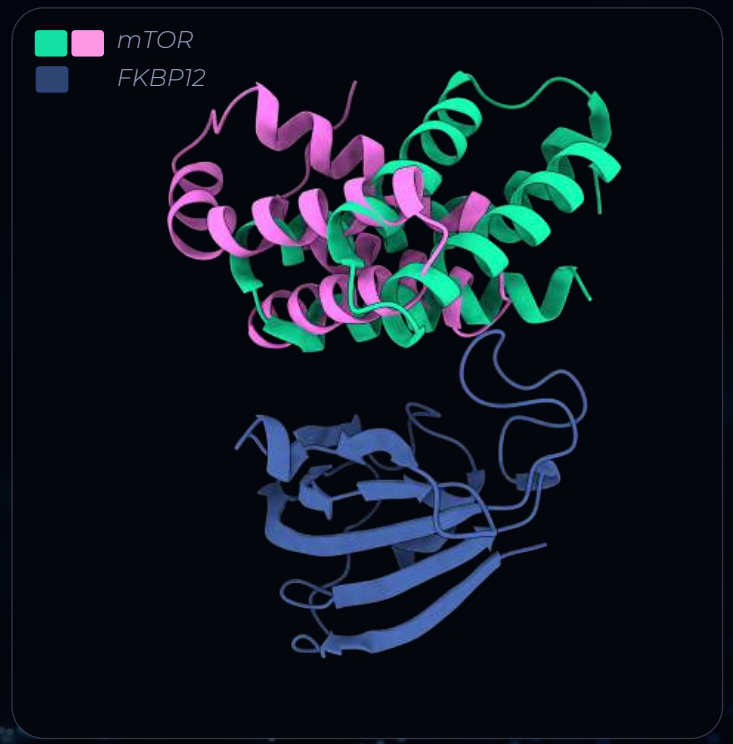


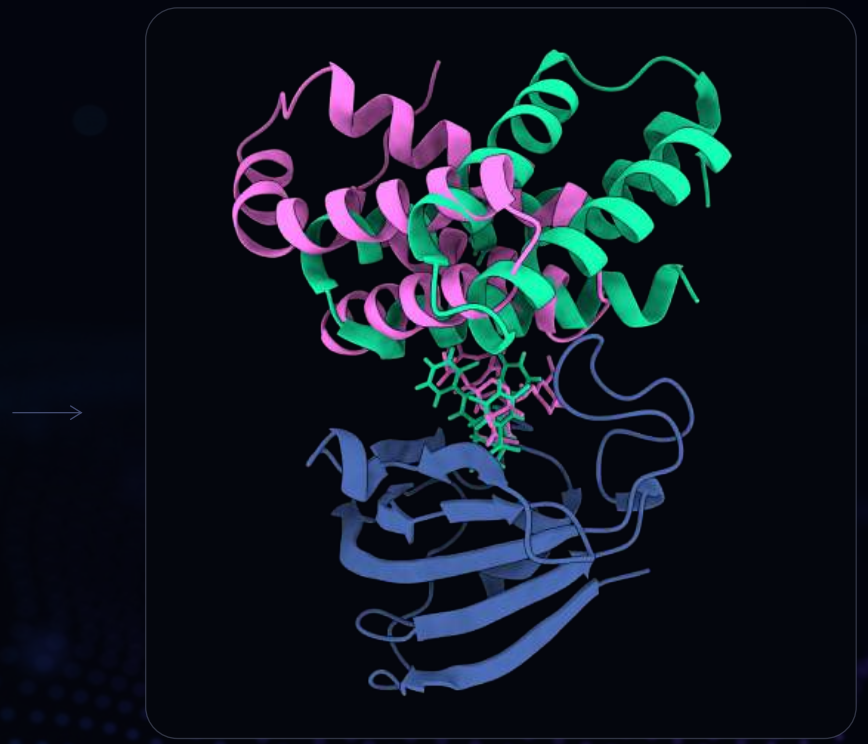
Both! Depending on molecule

6w8i & 6w70

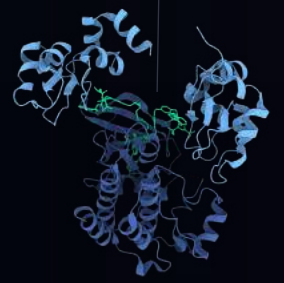
VANTAI

Which of these two FKBP12-mTOR Protein-interfaces is the right one?





Molecule-induced interface



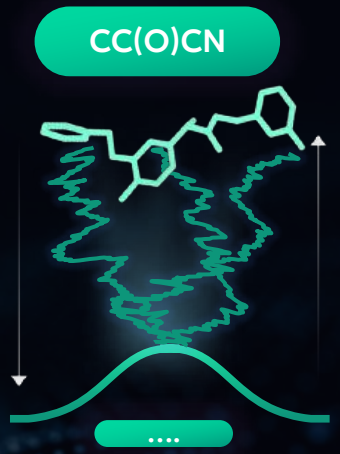
Both! Depending on molecule

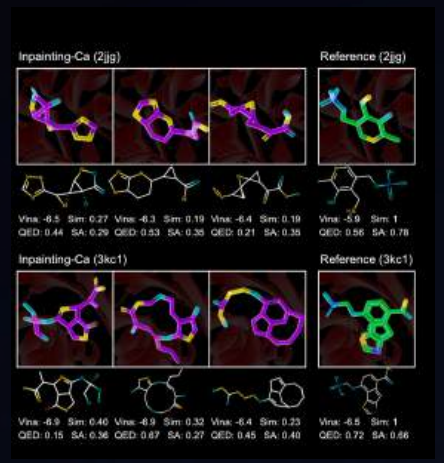
1FAP & 8PPZ

VANTAI

De-novo design:

De-novo small molecule design still lag behind:
a) methods assume knowledge of bound-state protein & b) struggle with valid designs





Structure-based Drug Design with Equivariant Diffusion Models. Schneuing, ..., Bronstein, Correia. 2022

VANTAI

Introducing

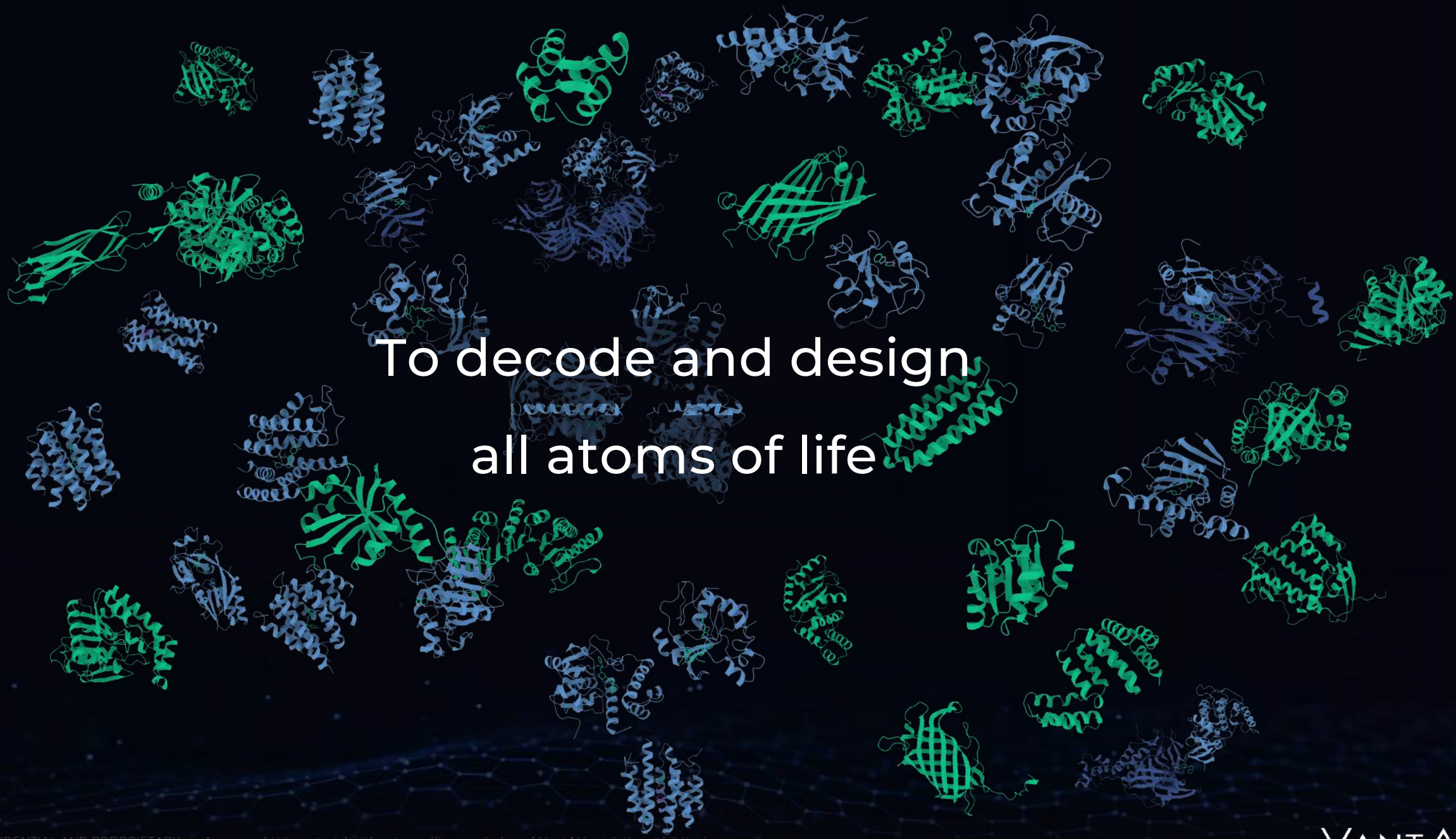
VantAI

NEO-1

vant.ai/neo-1

Introducing Neo-1: the worlds most advanced atomistic foundation model

VantAI



VANTAI