Lab in the loop Machine learning for LMDD

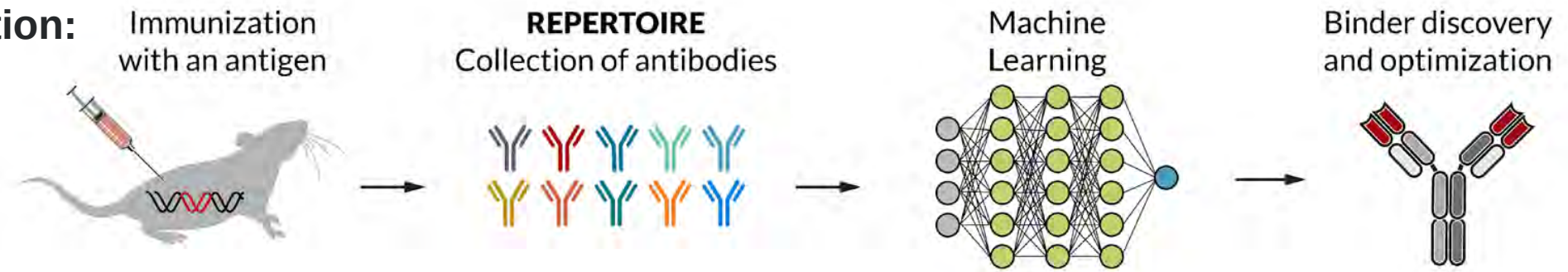
Richard Bonneau

Prescient Design
A Genentech Accelerator

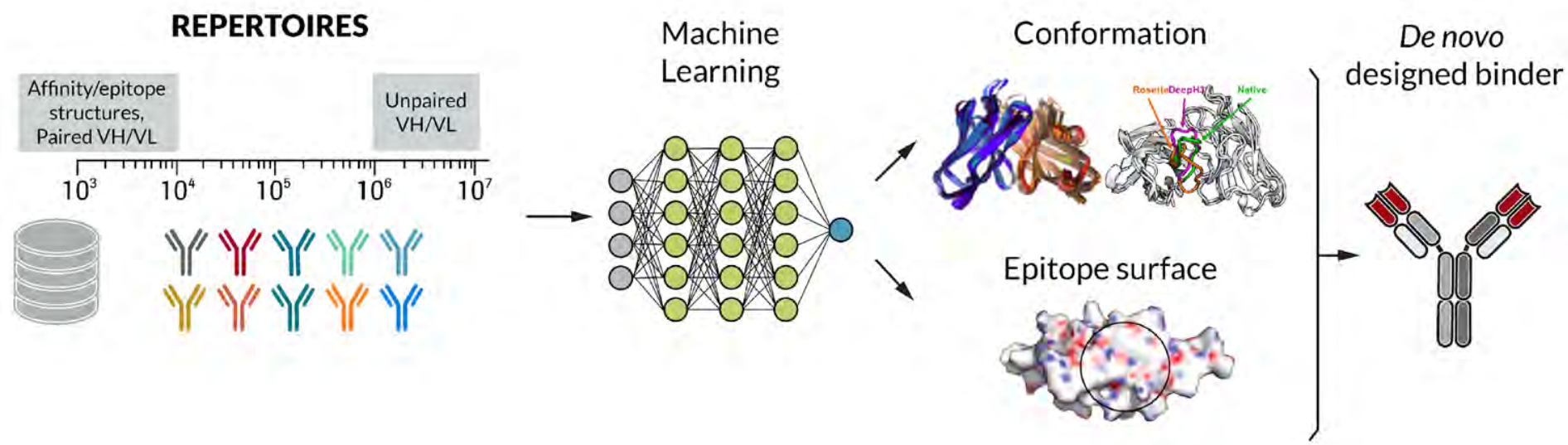
Genentech
A Member of the Roche Group

AI / ML for antibody identification, optimization and *de novo* design

Optimisation:

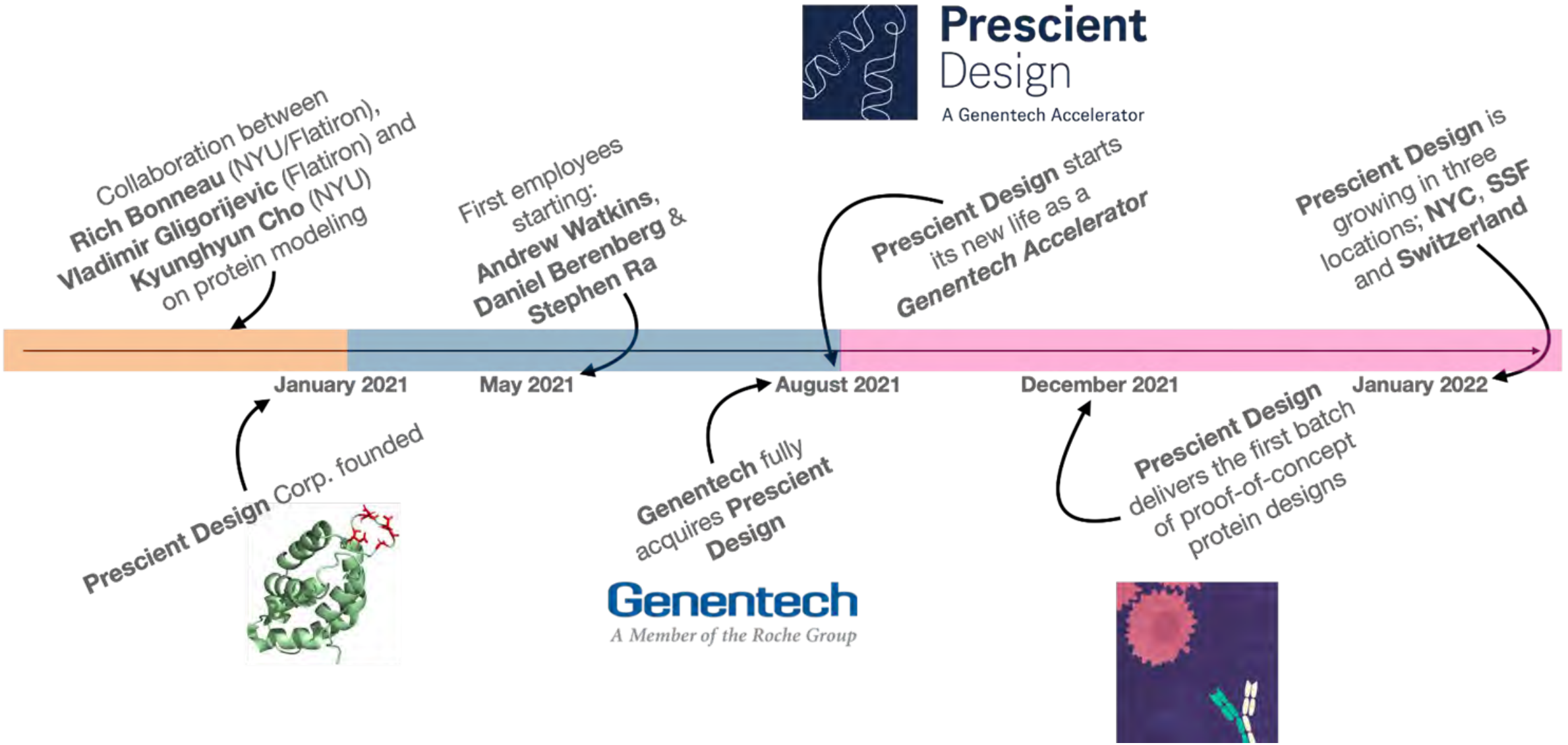


Direct:



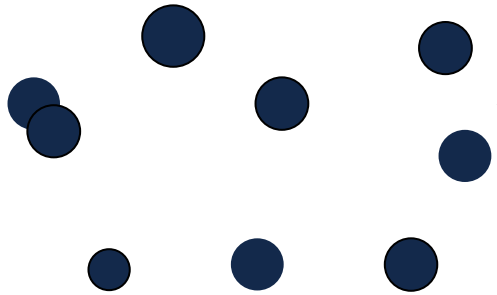
Ruffolo et al. Geometric potentials from deep learning improve prediction of CDR H3 loop structures. *Bioinformatics* **36** (1-Suppl.), i268-i275 (2020)
Gainza et al., Protein interaction fingerprinting using deep learning. *Nature Methods* **17** *(2)*, 2020

Brief History of Prescient Design

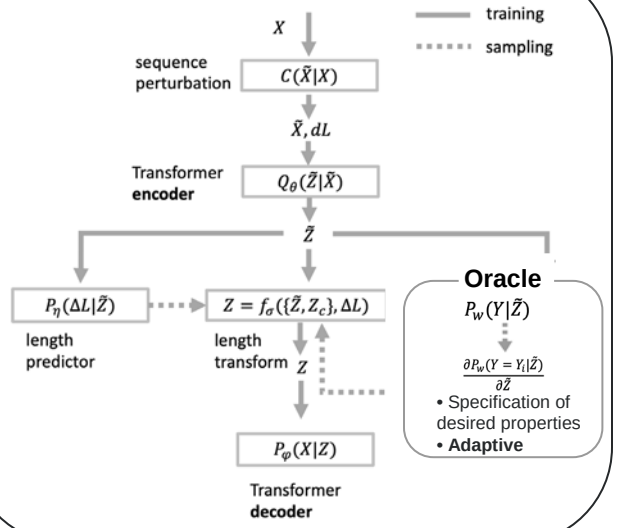


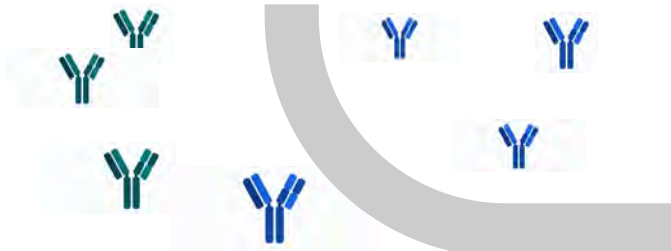
Lab-in-the-Loop Framework

An iterative framework for evidence acquisition, guided design, and validation



Manifold Sampler



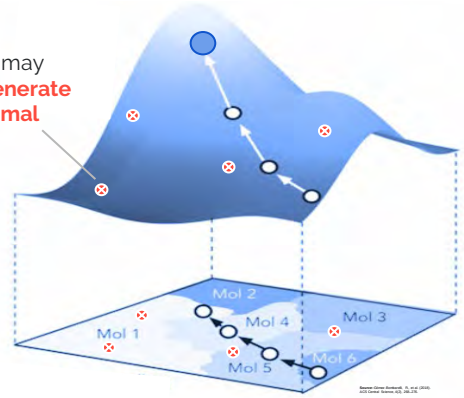


Generated designs with predicted properties *Y*

Refining the generative process recursively
The acquisition of experimental results -- both **positive and negative** -- combined with new data helps **adapt** the generative process and enables better optimization of future designs

Guided sampling

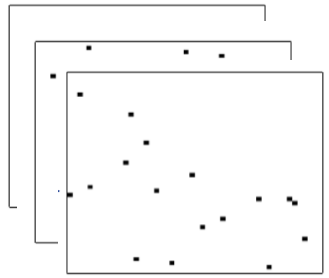
Random mutations may yield **degenerate** or **suboptimal** designs



- Gradient-based **multi-objective** guidance
- Active learning & **Uncertainty-awareness**
- **Efficient sampling** of design space

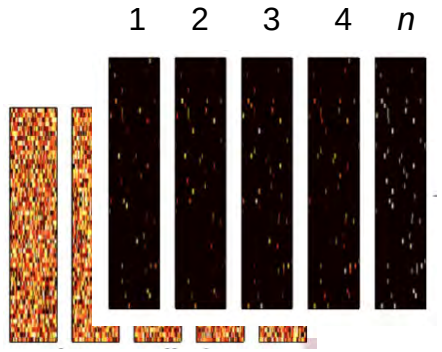
Proposing new designs
Our sampling procedure finds an **optimal exploration-exploitation strategy** and can query SMEs for experimental validation/synthesis of generated designs

Experimental feedback



- Selection of optimized leads
- Evaluation data for accepted/rejected proposals
- Additional SME feedback

Assay

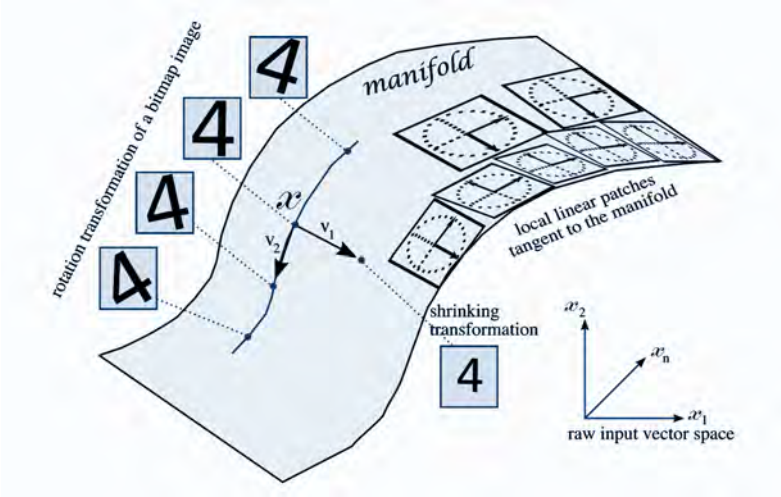


Experimental validation of proposed designs

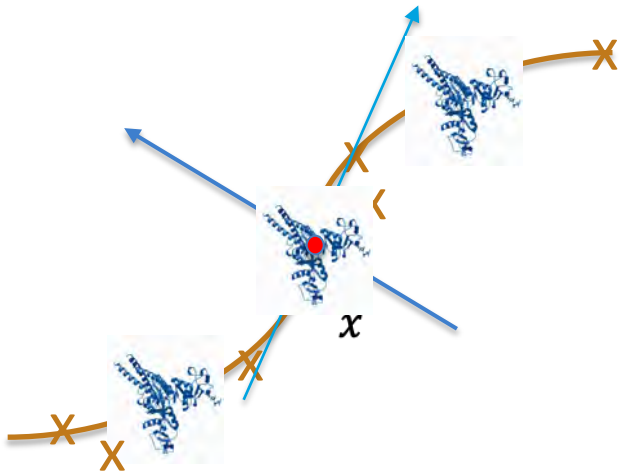
Manifold Sampler

Vladimir Gligorijevic, Dan Berenberg, Stephen Ra, Andy Watkins, Simon Kelow, Rich Bonneau, Kyunghyun Cho

Real-world, high-dimensional data lie (roughly) on a **low-dimensional manifold** (Chapelle 2006)



Bengio (2009)



Instead of the (**large**) combinatorial space, search for a novel sequence x is restricted on the (**small**) manifold space, i.e., $x \in \mathcal{M}$

Function-guided design: $\underset{\mathbf{x} \in \mathcal{M}}{\operatorname{argmax}} f(\mathbf{x})$ for a scoring function $f(\mathbf{x})$ trained on $\mathcal{M}$ where $f(\mathbf{x})$ is an oracle.

Given a protein (target) sequence $Y = y_1, \ldots, y_n$, **learn a Seq2Seq DAE** model where X is a corrupted version of Y: $X \rightarrow Y$

$$Y \quad \mathbf{X} \sim C(\mathbf{X}|\mathbf{Y}) \qquad X$$

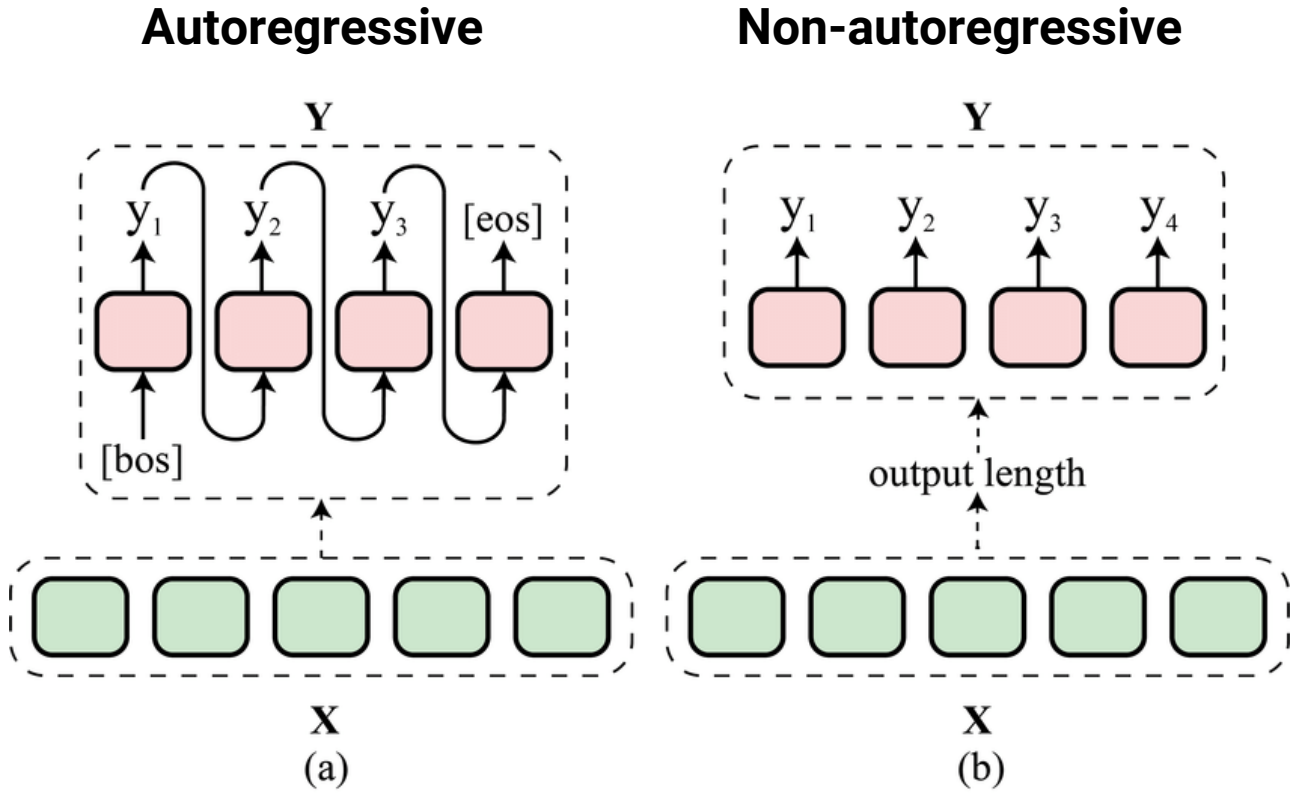
Two approaches:

1. Autoregressive Seq2Seq:

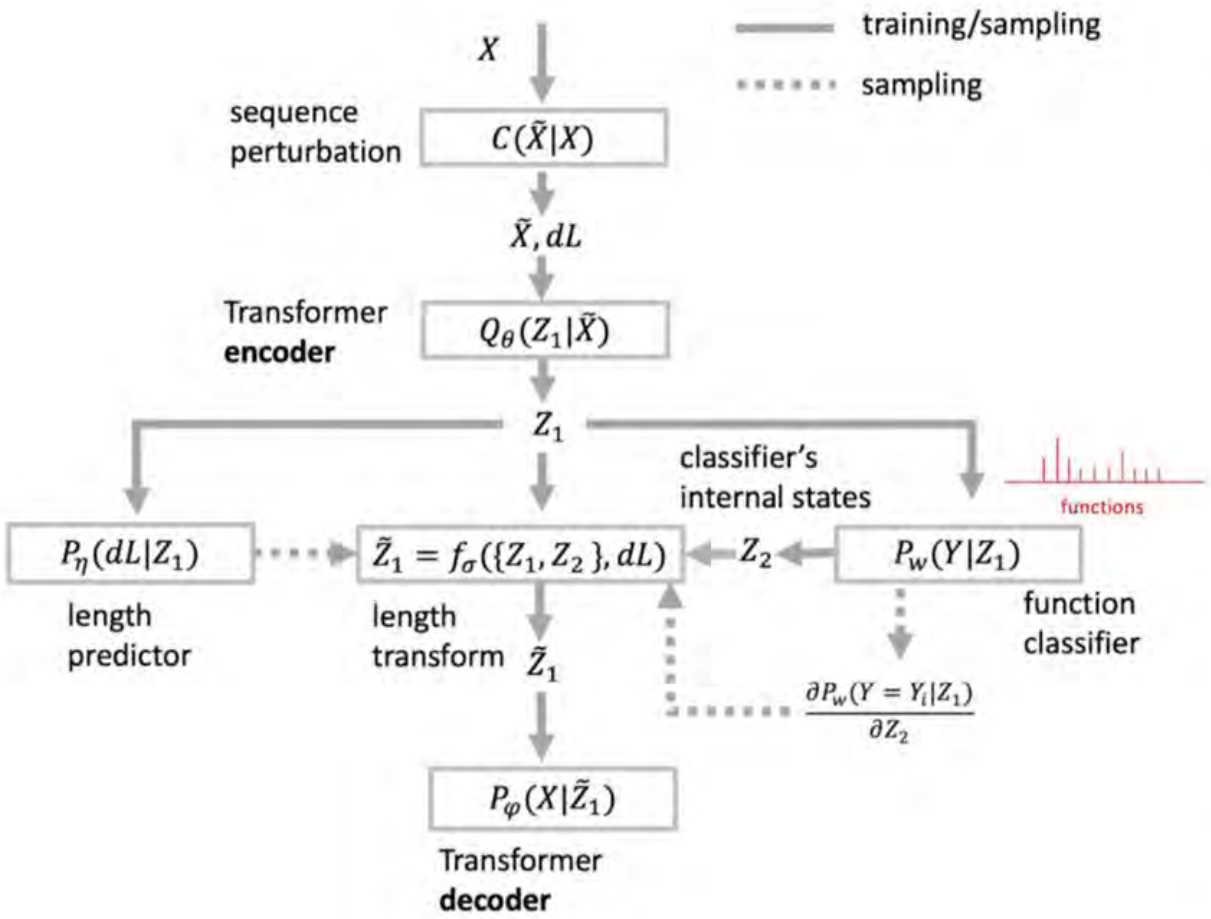
1. Non-autoregressive Seq2Seq:

$$\log p(\mathbf{Y}|\mathbf{X}) = \sum_t \log p(y_t, y_{<t}, \mathbf{X})$$

$$\log p(\mathbf{Y}|\mathbf{X}) = \sum_t \log p(y_t, \mathbf{X})$$



Non-autoregressive sampling. Makes changes in multiple positions of a target sequence enabling **effective** exploration of the overall fitness landscape and resulting in **diverse** sequence designs



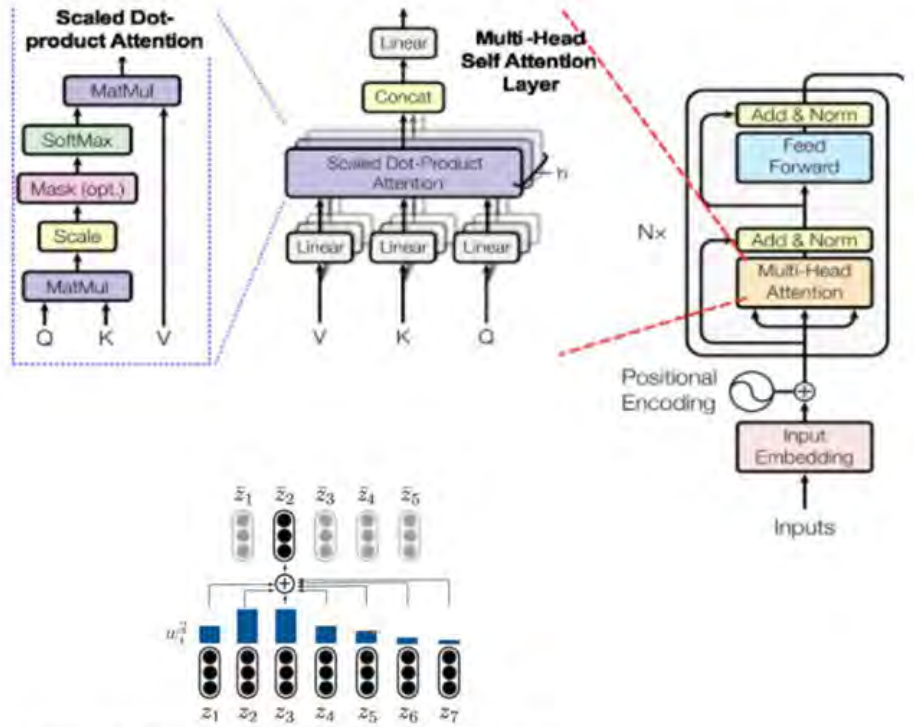


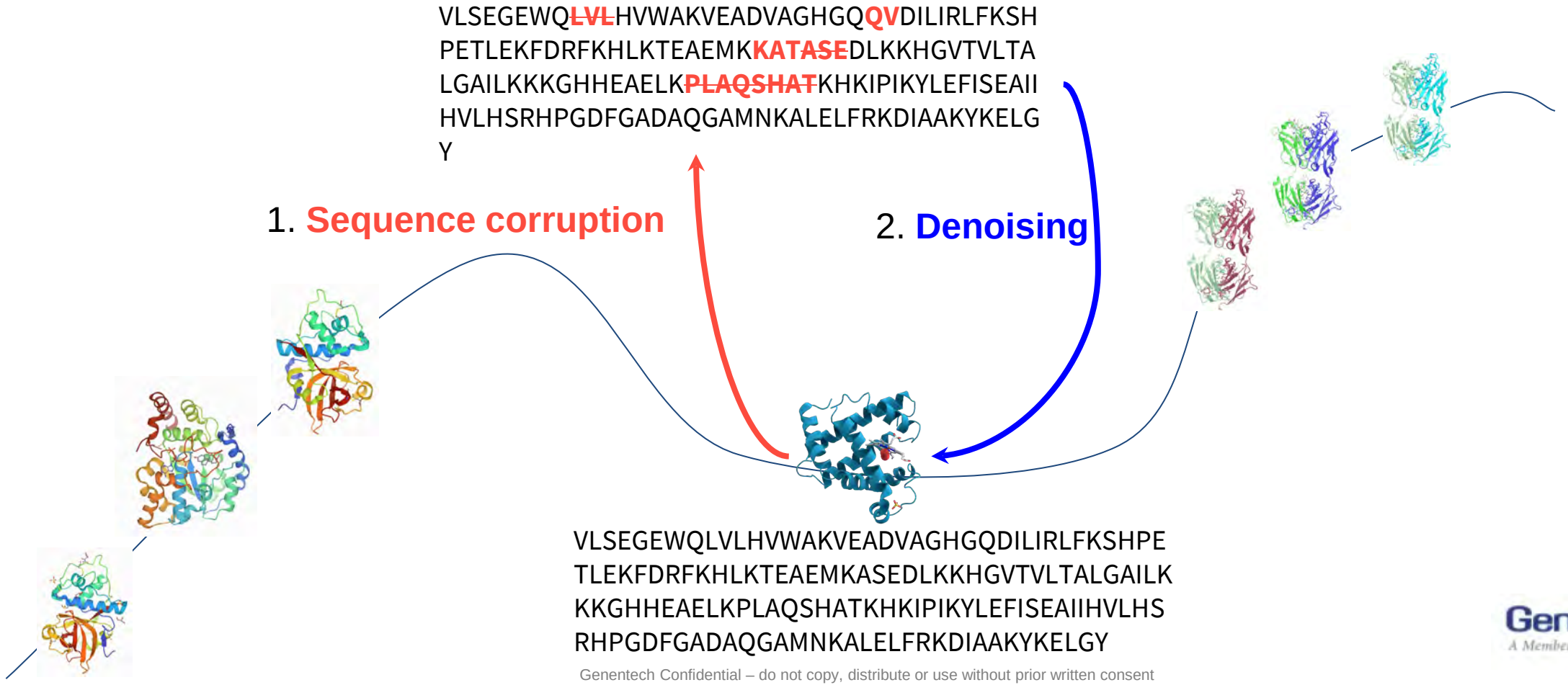
Figure 2: Illustration of the length transformation mechanism.

Shu, Raphael, ..K. Cho "Latent-variable non-autoregressive neural machine translation with deterministic inference using a delta posterior." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 05. 2020.

Antibody Design: Deep manifold sampler

Deep manifold sampler identifies the protein manifold by learning to denoise [Gligorijevic et al., 2021]

It thereby trades off between **statistical efficiency** and **combinatorial complexity**.



Applications to LMDD

More details in our recent publications

Article | Open Access | Published: 26 May 2021

Structure-based protein function prediction using graph convolutional networks

Vladimir Gligorijević ✉, P. Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C. Taylor, Ian M. Fisk, Hera Vlamakis, Ramnik J. Xavier, Rob Knight, Kyunghyun Cho & Richard Bonneau ✉

TM-Vec: template modeling vectors for fast homology detection and alignment

Tymor Hamamsy[1,*], James T. Morton[2,3,*], Daniel Berenberg[4,9], Nicholas Carriero[5], Vladimir Gligorijevic[9], Robert Blackwell[5], Charlie E. M. Strauss[7], Julia Koehler Leman[2], Kyunghyun Cho[1,4,8,9,‡], and Richard Bonneau[1,4,6,9,‡]

Function-guided protein design by deep manifold sampling

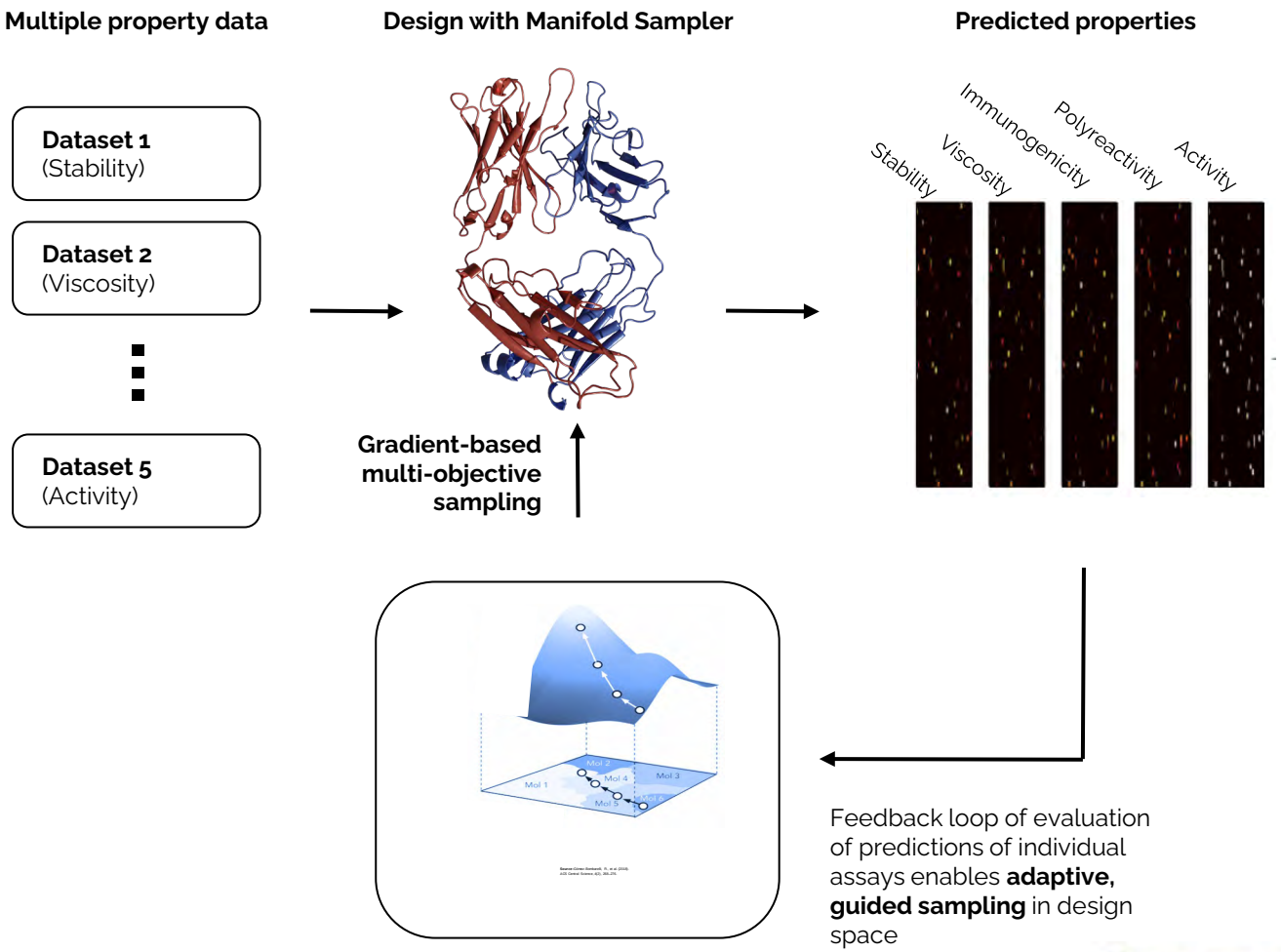
Vladimir Gligorijević[1], Daniel Berenberg[1,2], Stephen Ra[1], Andrew Watkins[1], Simon Kelow[1], Kyunghyun Cho[1,2,3], and Richard Bonneau[1]

Multi-segment preserving sampling for deep manifold sampler

Daniel Berenberg[1,2], Jae Hyeon Lee[1], Simon Kelow[1], Ji Won Park[1], Andrew Watkins[1], Vladimir Gligorijević[1], Richard Bonneau[1], Stephen Ra[1], and Kyunghyun Cho[1,2,3,4]

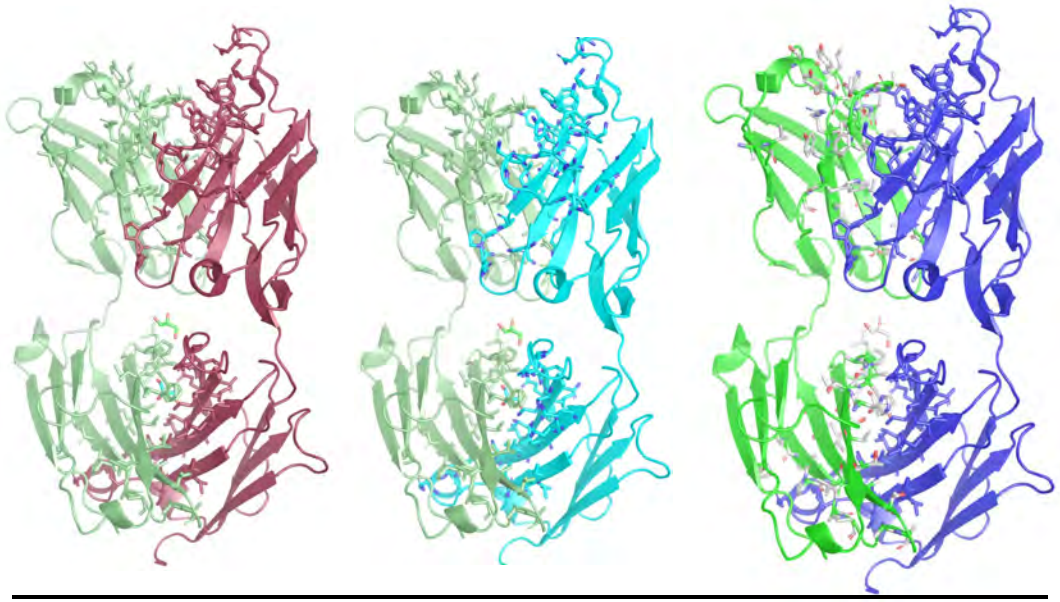
Multi-property optimization

- Parallel optimization of multiple properties
  - currently optimized sequentially
  - e.g. affinity, polyreactivity, viscosity
- Increased yield
  - increased number of candidate sequences that meet criteria
- Better optimisation
  - optimizing property *n* will not degrade property *n-1* as it might with sequential optimisation
- Meta-learning multiple properties
  - robustness for small training datasets
- Manifold Sampler naturally models epistasis

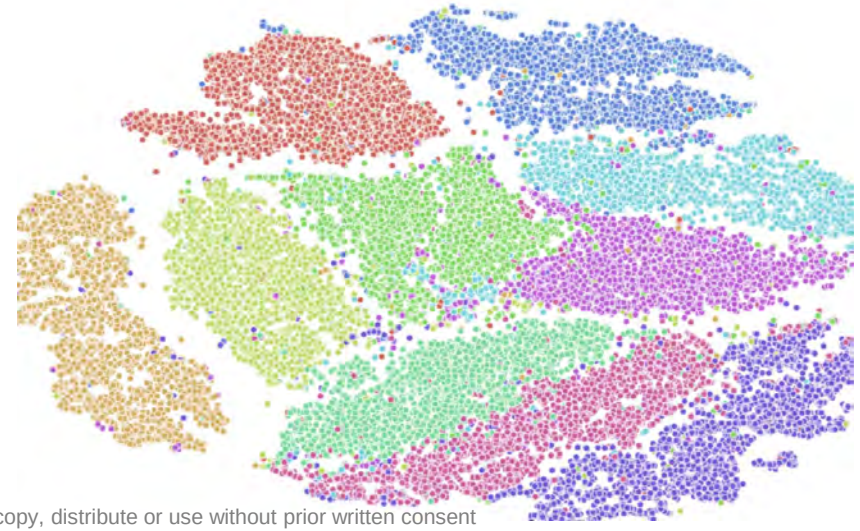


Chain pairing and epitope diversification

- Sufficient data on VH-VL pairing across multiple immunizations
- High quality negative examples (important for our approach) can be generated by pairing VH and VL across immunizations to disparate targets.
- Chain pairing will feed into better epitope clustering and identification of rare epitopes.
- Sorting paratopes into clusters corresponding to epitopes will be done in manifold representation.
- Representatives from clusters or rare epitopes feed directly into design

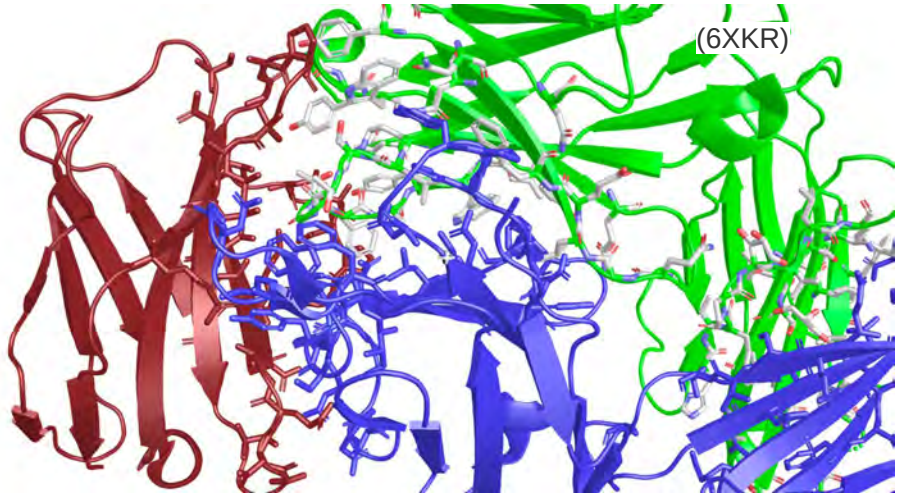


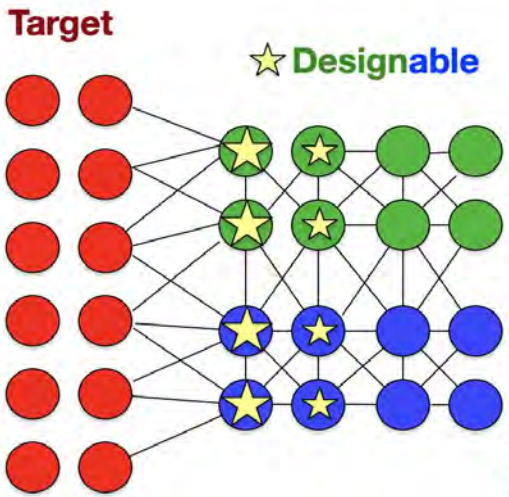
Paratopes clustered on manifold



De novo affinity design

- Given any target protein or interface the problem is well constrained.
- We first set up a graph representing the desired interaction to enable geometric deep learning.
- The Manifold Sampler is then used to power designs on that graph.
- Antibody (and protein-protein) interfaces are well represented in the PDB, and data to begin training this model is in-hand.
- Validation with structure determination and/or characterisation of binding feeds back on model (active learning).





Genentech
A Member of the Roche Group

Readouts from multiple iterations of lab-in-the-loop

Wide application:
2 development targets, Ova and Her. 6 ‘hot’ targets, 4 ‘medium’ targets
Applications to other Fab like formats. Application to non antibody affinity.

Active learning is working (not a disaster ;-)

Improvement in affinity and **multiple developability parameters** for 5 of 6 initial targets.

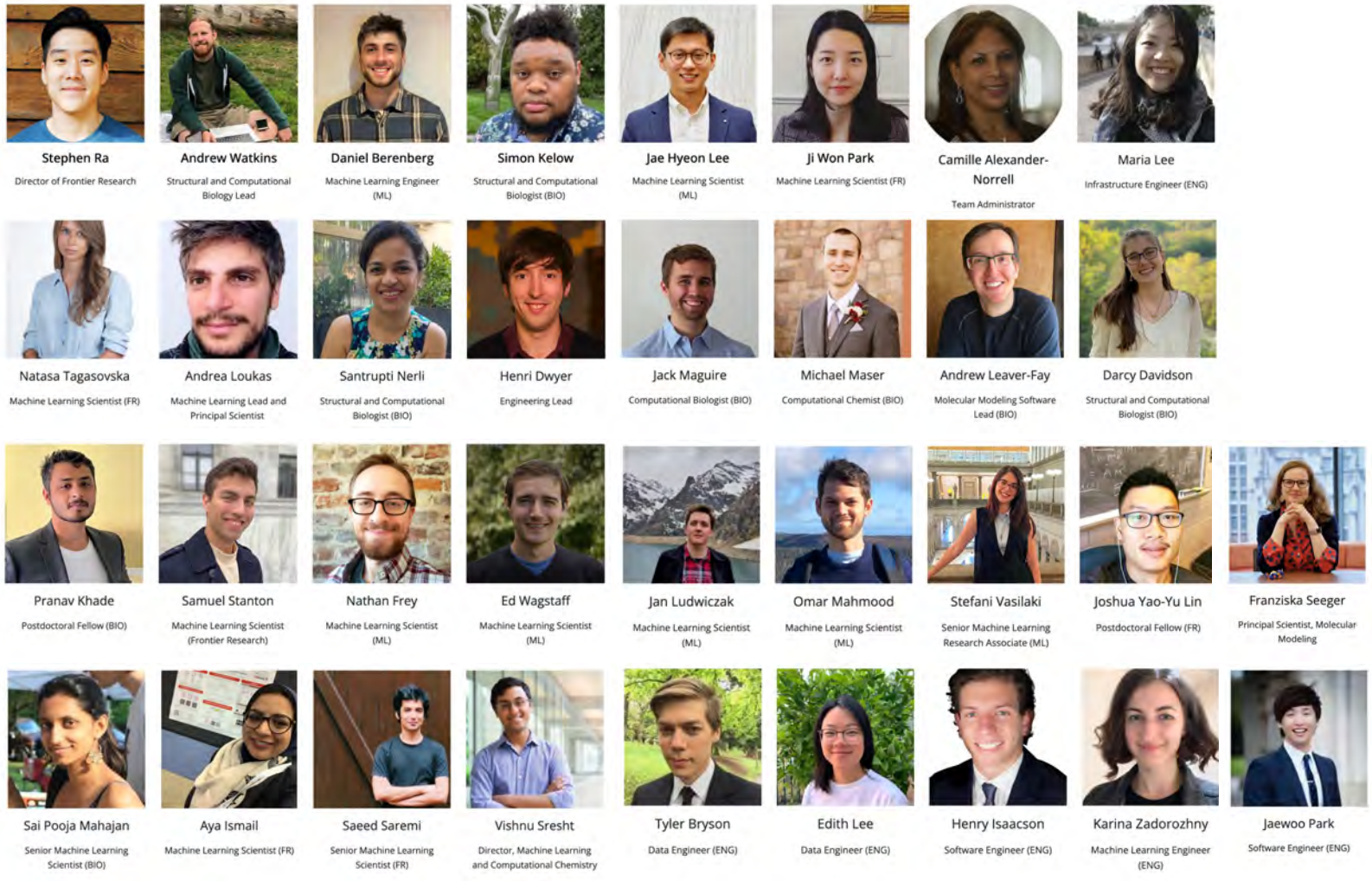
Mutations/Edit-distances of 5-65 … expression, binding and good dev for some molecules with > 50 mutations. True global design.

New high throughput lab techniques will drive even better approaches to active learning.

Good early readouts with ‘real’ *de novo* and grafting *de novo.* Much work remains.

Team

Prescient Design
A Genentech Accelerator



Even mix of:
- Structure Biology
- ML for DD
- Frontiers ML
- Engineering