Bayesian Optimization of Antibodies with a Generative Model of Evolving Sequences

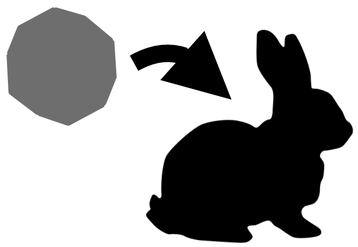
Alan N Amin, *Nate Gruver, *Yucen Lily L, *Yilun Kuang, Hunter Elliott, Calvin McCarter, Aniruddh Raghu, Peyton Greenside, Andrew G Wilson

NEW YORK UNIVERSITY

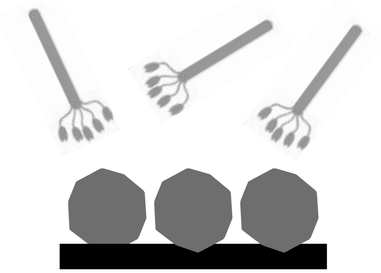
BigHat BIOSCIENCES

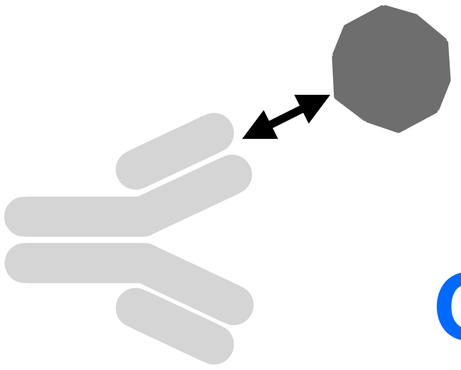
To build antibody drugs, we need to optimize “hits” to be strong binders that are stable in the human body

Animal models:



In vitro screens:

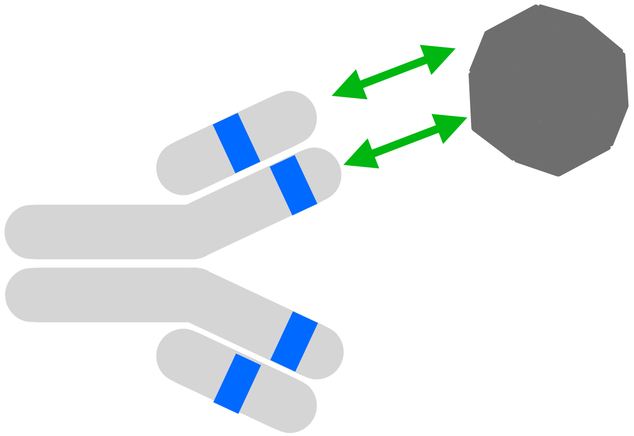




Starting candidate

Often:
Low affinity
Polyreactive
Low T_m

Optimize



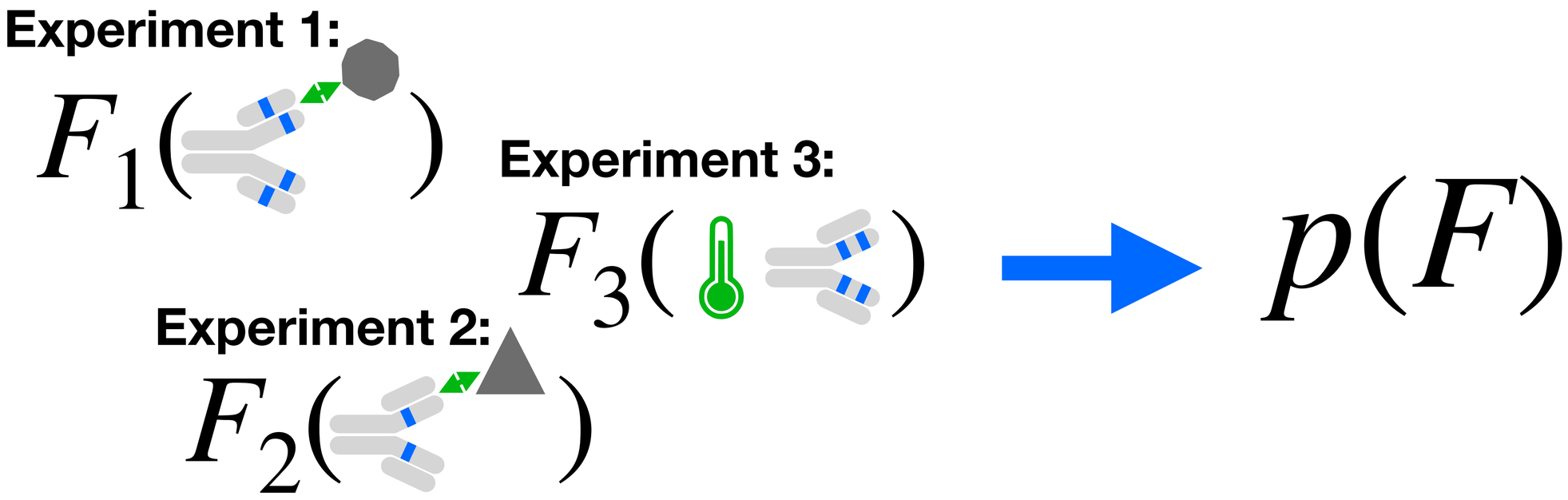
Optimization by iterative design is hard because most of the many mutations we can test do not help

X_0

QVQLRESGPGLVKPSQTLSLTCTVSGGSFNSGGYYWNRIRQHPGNGLEWIGYMYYSGSTYYNPFIRSRVIISGDTSVNHFSLKLSSVTAADTAVYFCARGYRQSGYSSSWVVDYWGQGTLVNVSS

- **Goal:** iteratively suggest and measure $X_1, X_2, \ldots, X_{100}$ to have no immunogenicity, have high affinity, or have high melting temperature, $F(X_N)$
- **Possible strategies:** random mutations, avoid mutations that don't often appear in humans
- **Challenge:** search space is huge

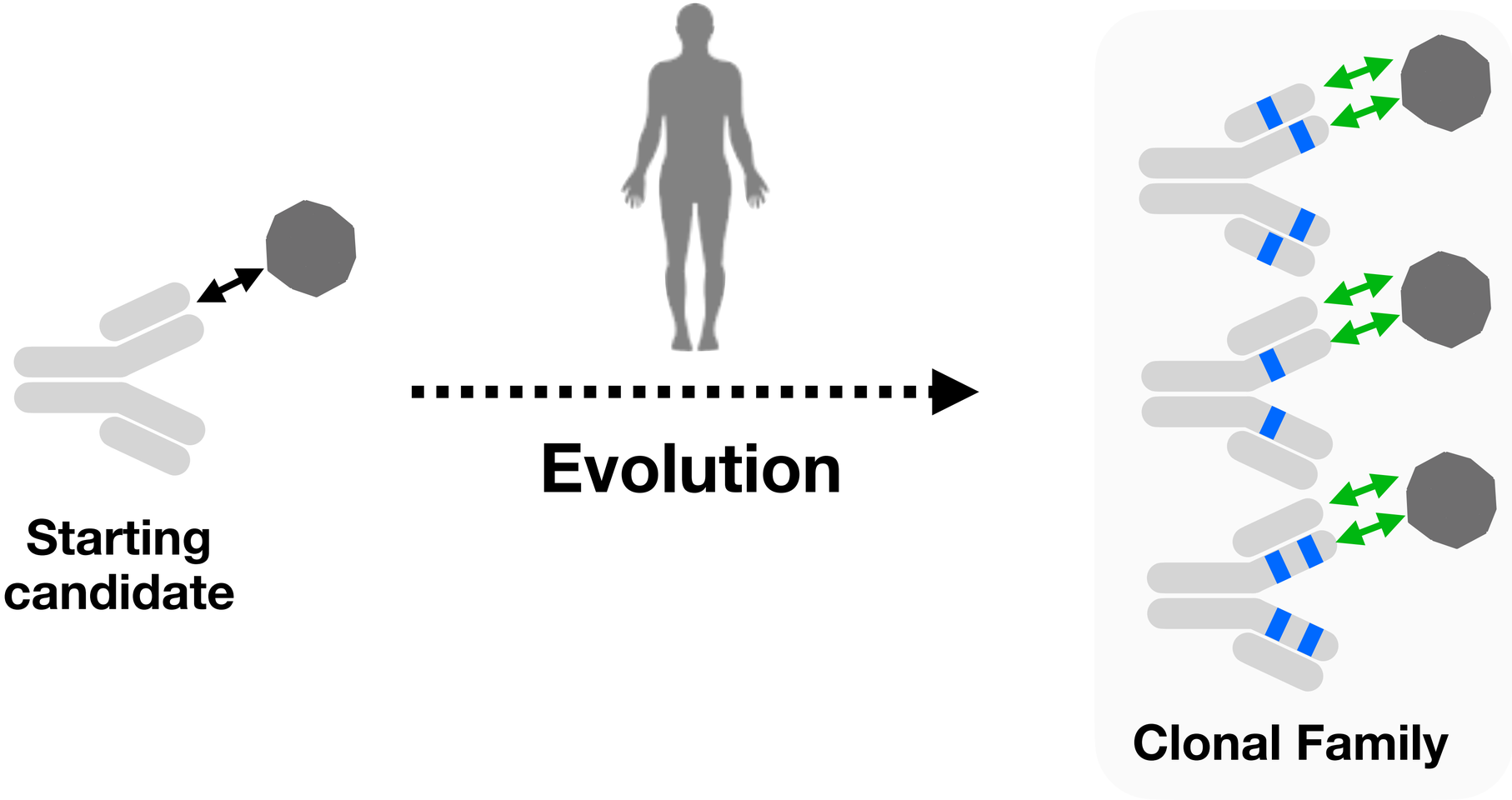
Ideally we could build a prior on the objectives $F(X)$ of binding strength and stability



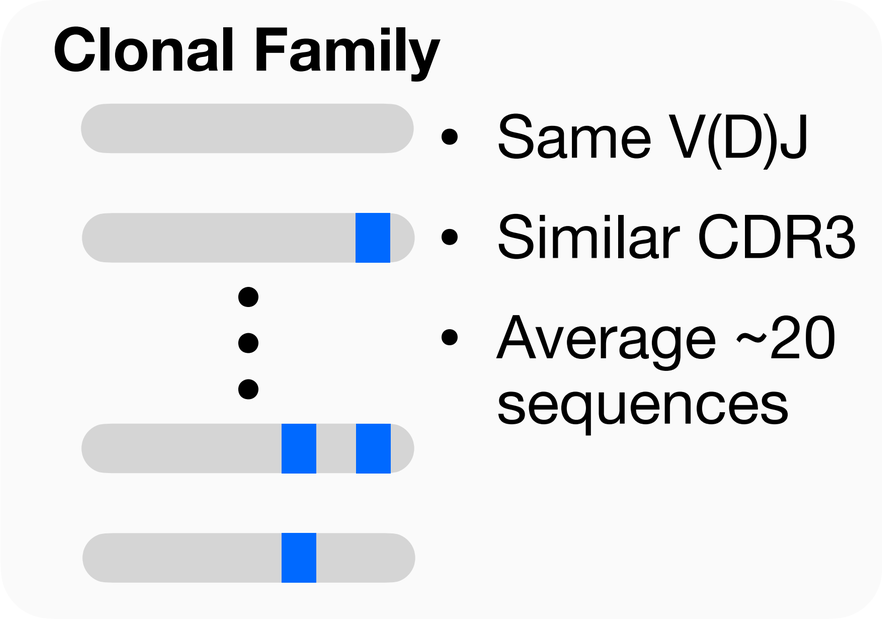
Optimal design strategy: suggest X_{N+1} based on $p(F \mid X_1, \ldots, X_N)$

But we don't have this data!

In principle, we can learn from how our body builds strong and stable binders



In principle, we can learn from massive data about human clonal families in the OAS database



Observed Antibody Space (OAS):

2×10^9 heavy chains

4×10^8 light chains

FastBCR
(Wang K et al 2023)

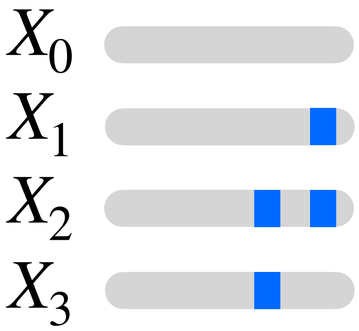
Observed Clonal Space:

9×10^5 heavy clonal families

3×10^4 light clonal families

In theory, we can build a prior over F by looking at the abundance of sequences in each clone

Distribution of sequences in a clonal family:

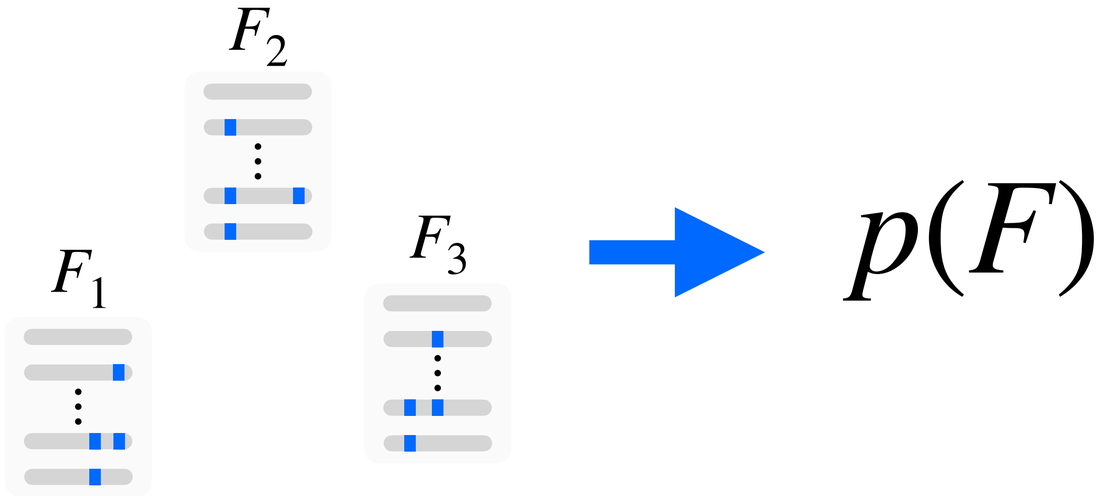


$$\sim_{\text{iid}} p(X \mid \text{clone})$$

Stronger, more stable binders are more abundant:

$$p(X \mid \text{clone}) = \text{Fitness}(X) =: F(X)$$

(Like protein families!)



CloneLM learns the distribution of clonal families

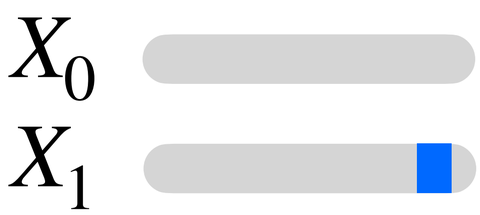
CloneLM (400 M transformer) trained on:

`seq1<separator>seq2<separator>seq3<separator>…`

Clonal family	QVQLRESGPGLVKPSQTLSLTCTVSGGSFNSGGYYWNRIRQHPGNGLEWIGYMYYSGSTYYNPFIRSRVIISGDTSVNHFSLKLSSVTAADTAVYFCARGYRQSGYSSSWVVDYWGQGTLVNVSS
	QVQLRESGPGLVKPSQTLSLTCTVSGGSINSGGYYWNWIRQHPGKGLEWIGYMYYSGSTYYNPFLRSRVIISADTSENHFSRKLSYVTAADTAVYFCARGYRQSGNSSSWVFDYWGQGTLVNVSS
	QVQLRESGPGLVKPSQTLSLTCTVSGGSINSGGYYWNWIRQHPGKGLEWIGYMYYSGSTYYNPFLRSRVIILADTSENHFSRKLSSVTAADTAVYFCARGYRQSGYSRSWVFDYWGQGTLVNVSS
	QVQLRESGPGLVKPSQTLSLTCTVSGGSINSGGYYWNWIRQHPGKGLEWIGYMYYSGSTYYNPFLSSRLIISADTPENHFSRRLSSVTAADTAVYFCATGYPQSGYSSSWVFNYWGQGTLVNVSS
Sample 1	QVQLQESGPRLVKPSQTLSLTCTVSGGSLNSGGYYWSWFRQPPGKRLEWIGYMYHTGNTYYNPSLKCRVTISGDTSKSHFPLRLTAVTAADTAAYYCARGYRQGGYSSSWLADYGGQGTLGADSS
	QVQLQESGPRLVKPSQTLSLTCTVSGGSLNSGGYYWGWIRQPPGKGLEWIGYMYHTGNTYYNPSLKSRVTISGDTSKNHFSLRLTSVTAADTAVYYCARGYRQGSYSSSWLADYWGQGTLVTVSS
	QVQLRESGPGLVKPSQTLSLTCTVSGGSFNSGGYYWNWIRQHPGKGLEWIGYMYYSGSTYYNPSLRSRVTISGGTSVNPFSLKLSSVTAADTAVYFCARGYRHSGYSSSLLVDYWAEETVVNVSS
Sample 2	QVQLRESGPGLVKPSQTLSLTCTVSGGSFNSGGYYWNWIRQHPGKGLEWIGYMYYSGSTYYNPYLRSRVIISGDTSENQFSLKLSSVTAADTAVYLCPRGYRQSCYSSSWVFDYWGQGTLVTVSS
	QVQLRESGPGLVKPSQTLSLTCTVSGGSFNSGGYYWNWIRQHPGKGLEWIGYMYYSGSTYYNPSLRSRVIISGDTSENHFSLKLSSVTAADTAVYFCARGYRQSGYSSSWVLDYWGQGTLVTVSS
	QVQLRESGPGLVKPSQTLSLTCTVSGGSFNSGGYYWNWIRQHPGKGLEWIGYMYYSGSTYYNPSLRSRVIISGDTSENHFSLKLSSVTAADTAVYFCARGYRQSGYSASWVFDYWGQGTLVTVSS
Sample 3	QVQLRESGPGLVKPSQTLSLTCTVSGGSFNSGGYYWNWIRQHPGNGLEWIGYMYYSGSTYYNPFLKSRVIISGDTSVTHFSLKLSSVTAADTAVYFCARGYRQSGSSSSWVIDYWGQGTLVTVSS
	QVQLRESGPGLVKPSQTLSLTCTVSGGSFNSGGYYWNWIRQHPGNGLEWIGYMYYSGSTYYNPFLMSRVIIRGETSVKHFSLKLRSVTAADTAVYFCARGYSQSGYSSSWVIDYWGQGTLVTVSS
	QVQLRESGPGLVKPSQTLSLTCTVSGGSFNSGGYYWNWIRQHQGDGLEWIGYLYYSGSTYYNPFVKRRVIISGDKSVNHFSLKLSSVTAADTDVYFCARGYGQSGYSSAWVIDYWGQGTLVTVSS

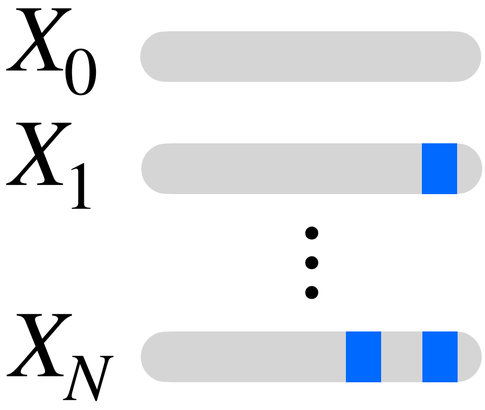
In theory, CloneLM performs approximate Bayesian inference over evolutionary landscapes

Predictions should integrate over F:



$$X_2 \sim \int p(X_2 \,|\, \text{clone}) dp(\text{clone} \,|\, X_0, X_1)$$

Predictions should converge to F:

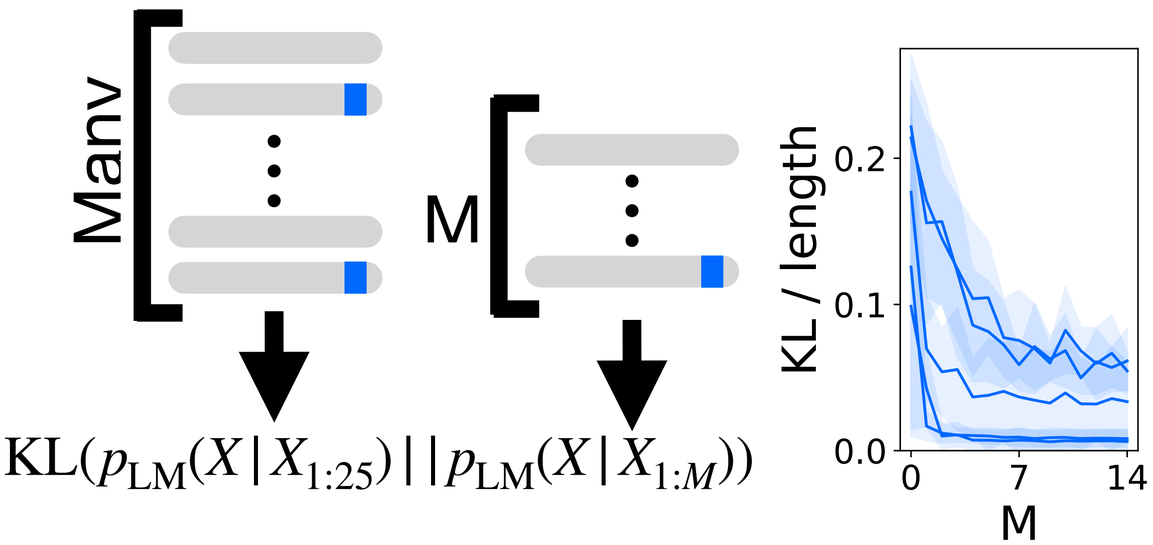


$$X_{N+1} \sim \int p(X_{N+1} \,|\, \text{clone}) dp(\text{clone} \,|\, X_{1:N})$$

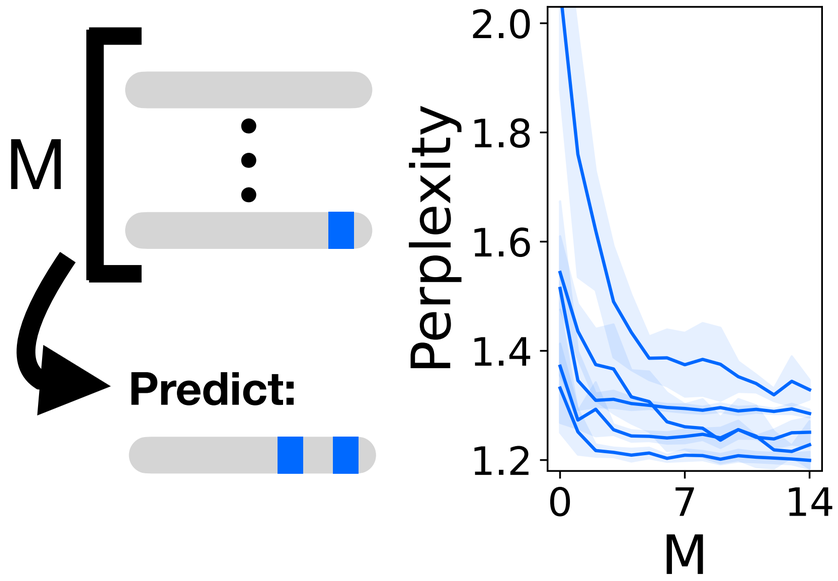
$$\approx p(X_{N+1} \,|\, \text{clone})$$

Given more sequences from a clonal family, CloneLM predictions converge to $p(X \mid \text{clone})$

Predictions converge:

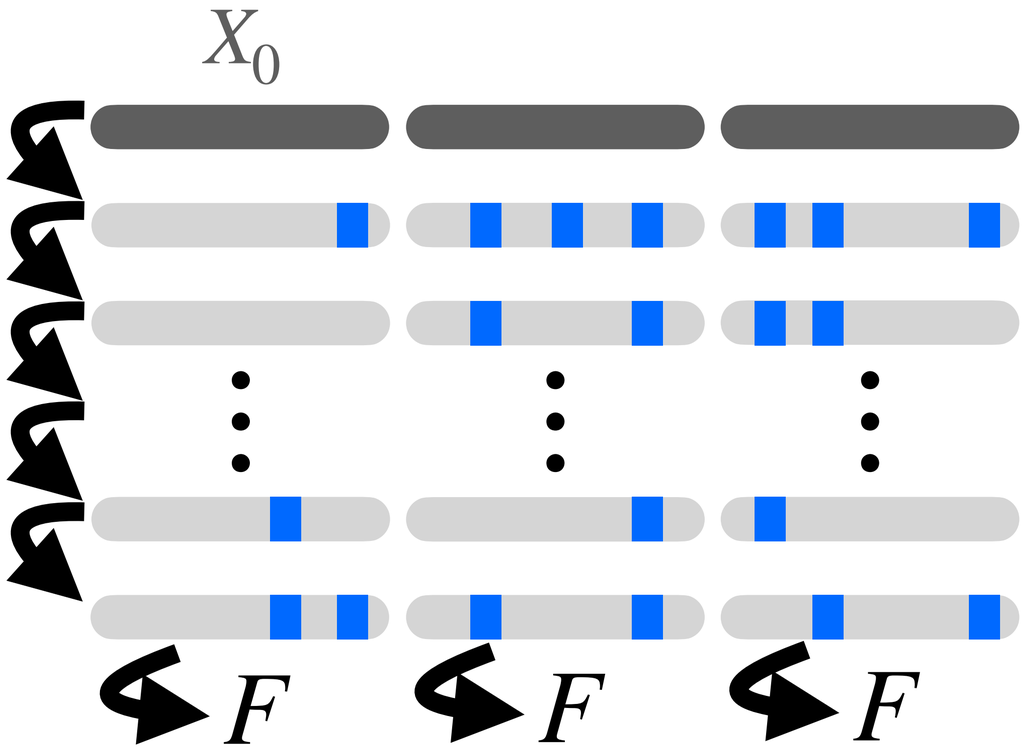


Predictions approach $p(X \mid \text{clone})$:

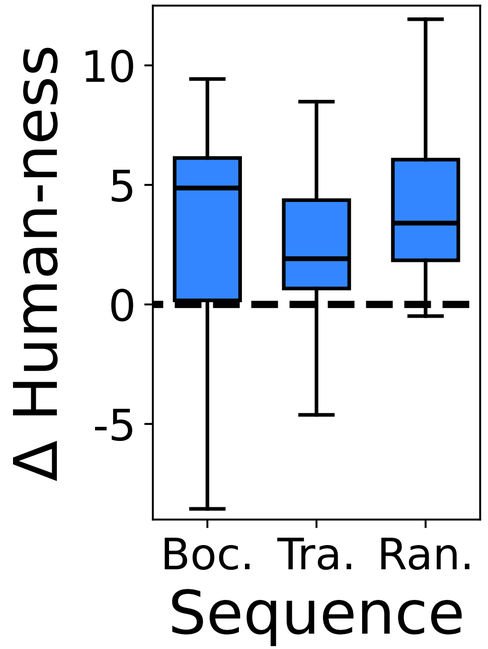


The CloneLM prior over fitness functions optimizes antibodies to become more human-like

Sample possible fitness landscapes F:

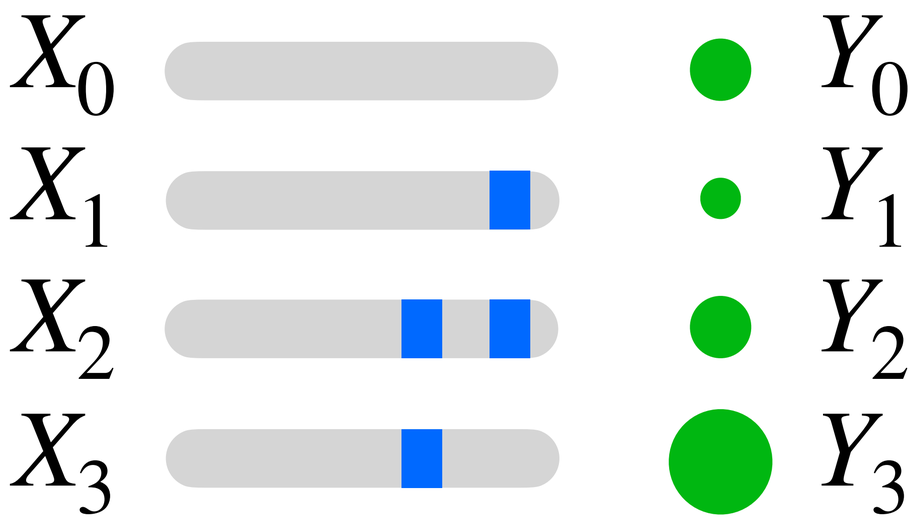


Optimize over three mutations with respect to F:



To update our belief in F, we assume measurements in the lab are proportional to fitness

Experiment:



Posterior:

$$p(F \mid X_0, (X_n, Y_n)_n) \propto p(F \mid X_0) \prod_n p(Y_n \mid F(X_n))$$

Likelihood:

$$p(Y \mid F(X)) \sim \mathcal{N}(\beta F(X) + C, \sigma^2)$$

Uniform prior on β, C

How do we sample from the posterior?

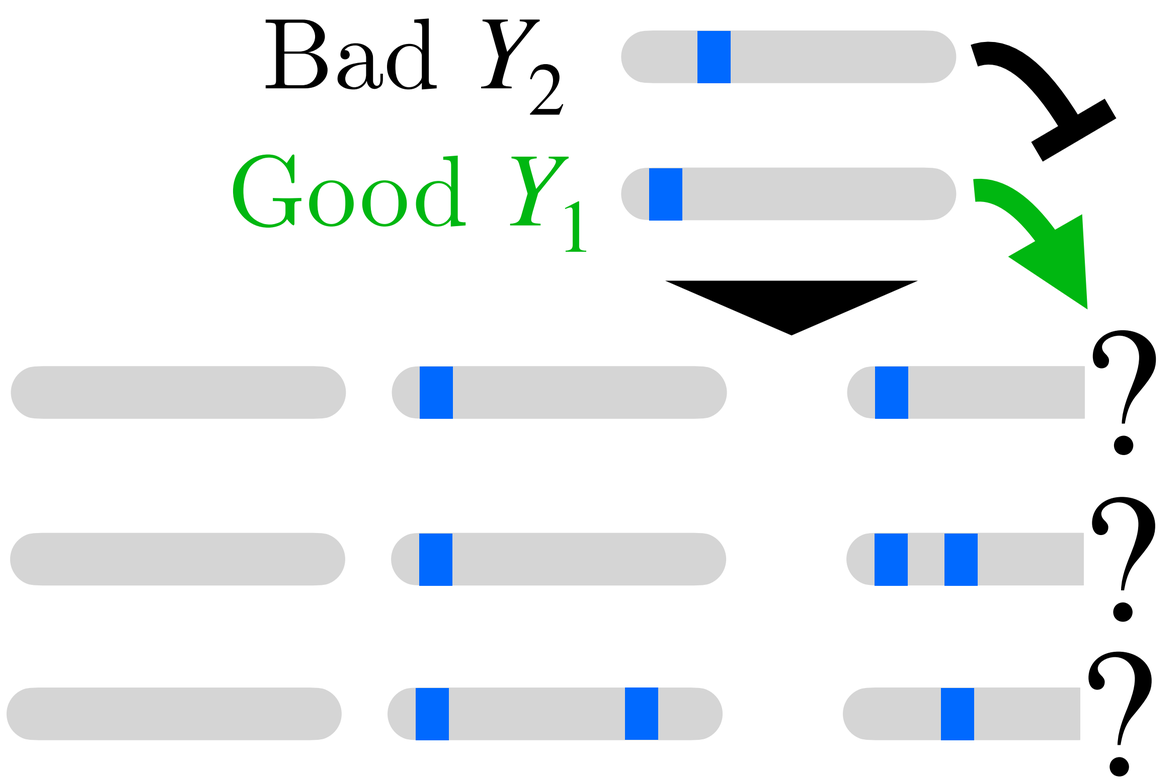
Importance sample (naive):

1. Sample many F from prior $F \sim p(F \mid X_0)$

2. Resample based on likelihood

$$\prod_n p(Y_n \mid F(X_n))$$

We inform our sampling with twisted stochastic Monte Carlo

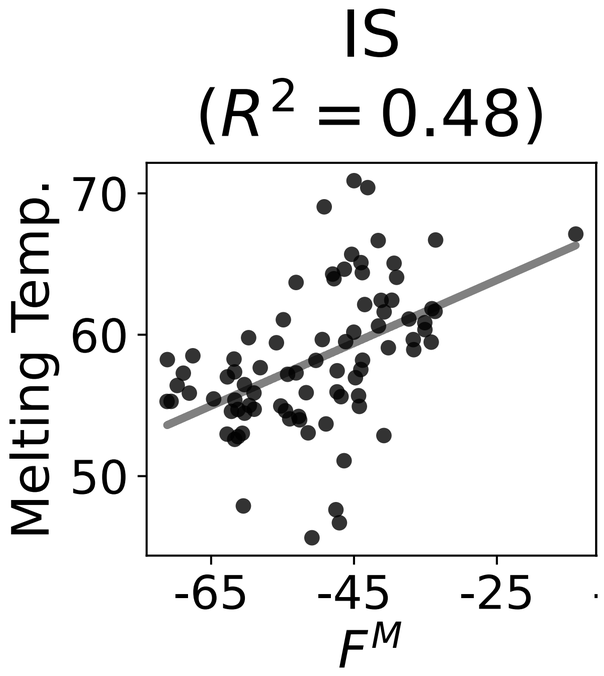


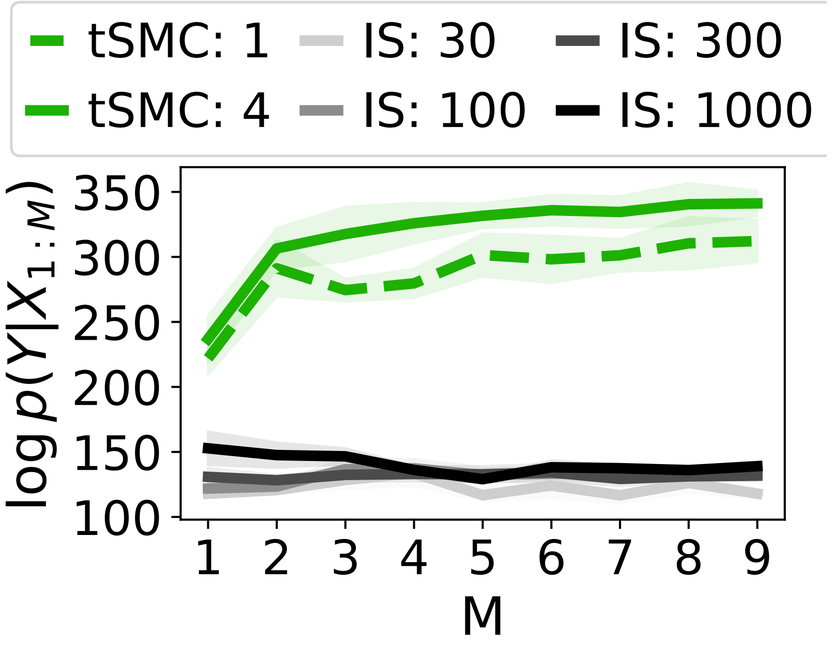
1. Sample letter based on heuristic

2. Resample after generating each letter to eliminate bias

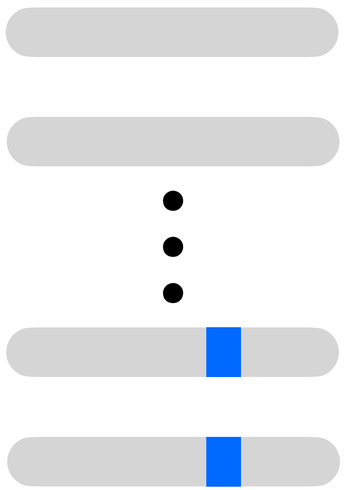
tSMC efficiently samples from the posterior

Condition on real T_m measurements



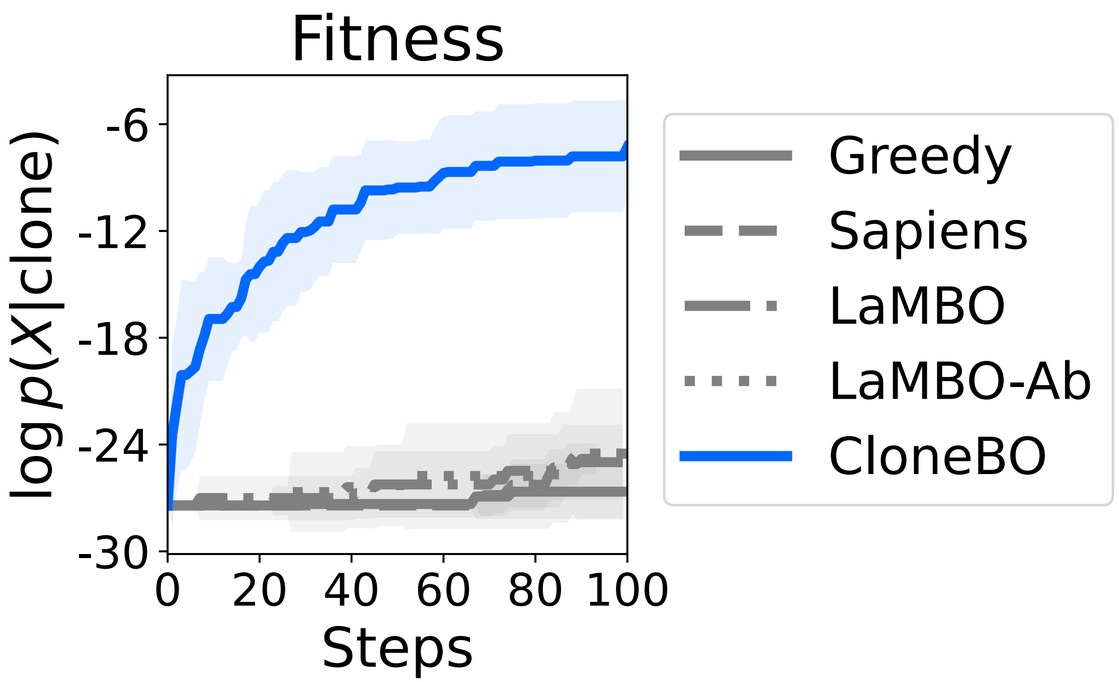


CloneBO efficiently optimizes a function from its prior



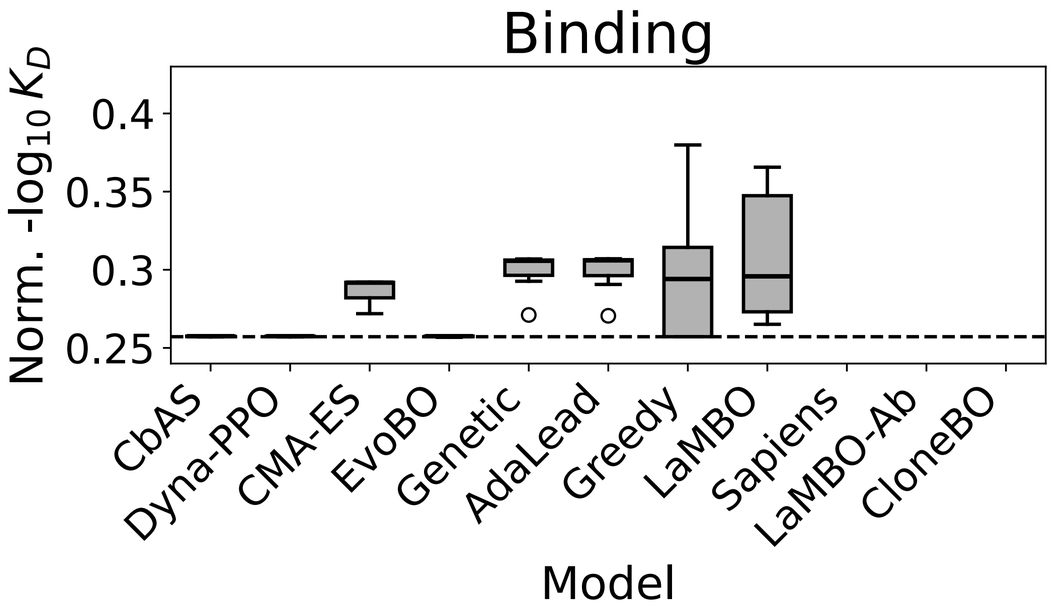
Train an LLM on a clone

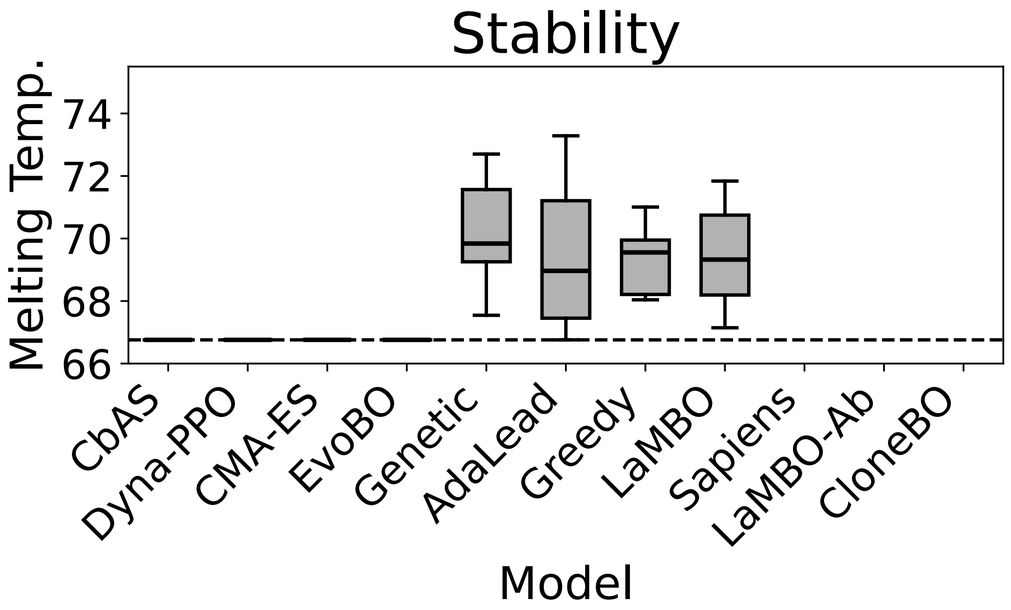
$$F(X) = \log p_{\mathrm{LLM}}(X \mid \mathrm{clone})$$



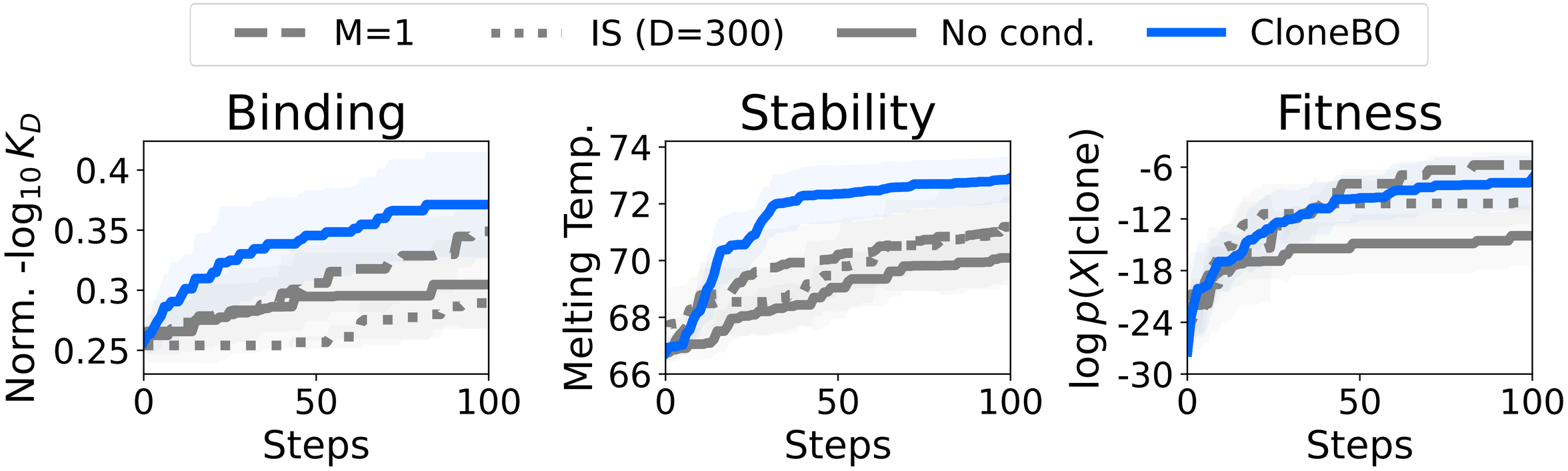
CloneBO efficiently optimizes for binding and stability *in silico*

$F(X) =$ neural network trained on thousands of sequences from iterative design experiment





Ablations demonstrate that CloneBO efficiently optimizes sequences by doing accurate inference

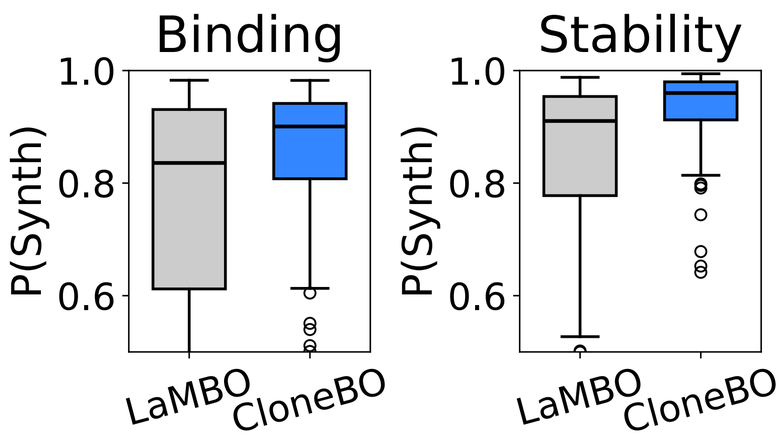


(Posterior is not hard to approximate in this case)

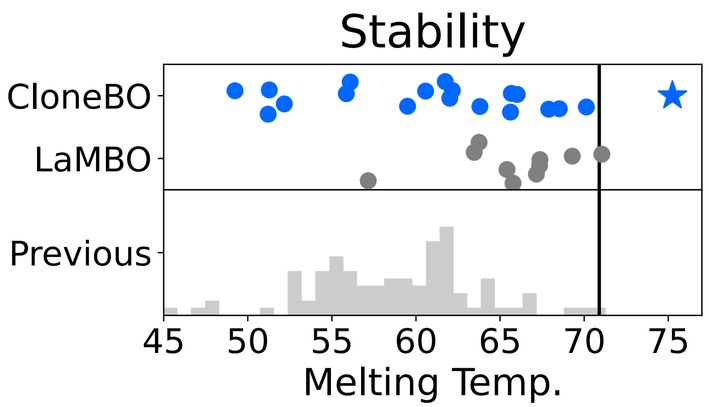
CloneBO efficiently optimizes for binding and stability *in vitro*

Given $X_0, \ldots, X_{1000}$, measurements for binding and stability, design X_{1001}

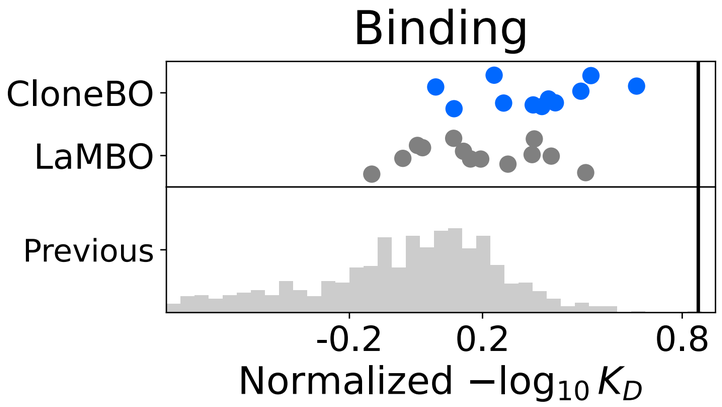
Designs are predicted to express:



Designs improve stability:

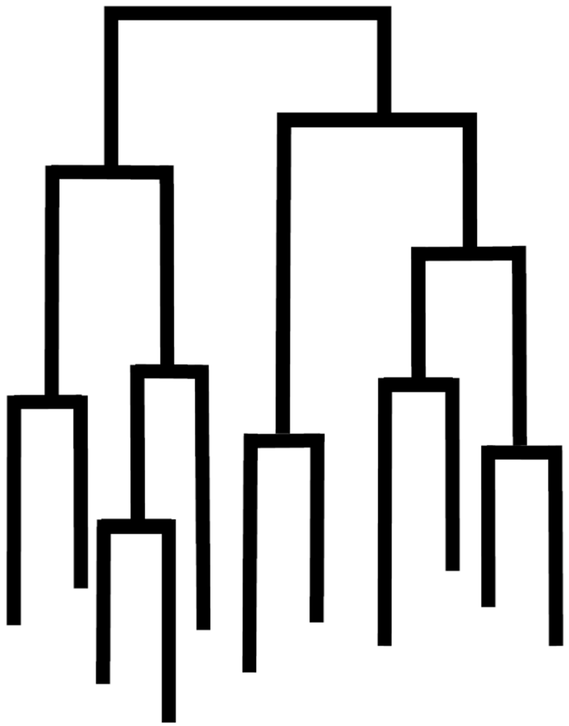


Designs outperform in binding:

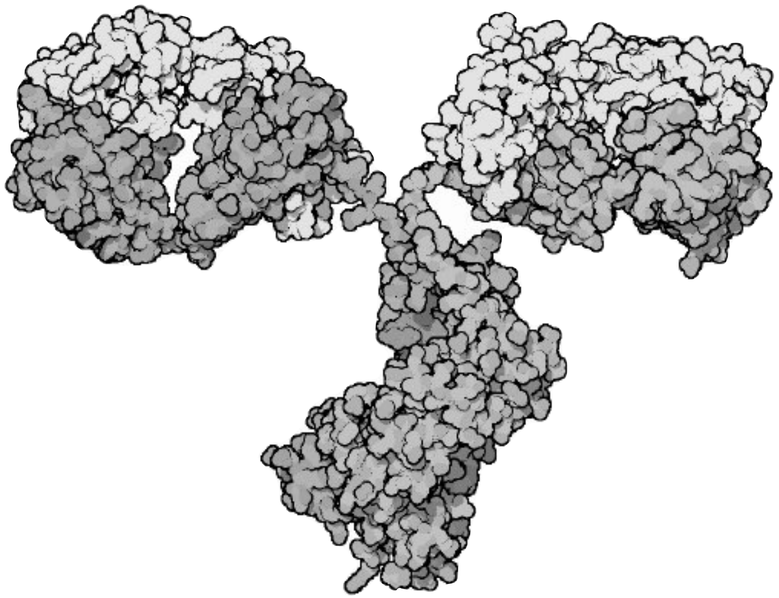


Future directions

More realistic prior
Learn direction of evolution



Build a prior on structure
Structure from clonal families



Acknowledgments

NEW YORK UNIVERSITY

BigHat BIOSCIENCES

Nate Gruver

Lily Li

Calvin McCarter

Aniruddh Raghu

Yilun Kuang

Andrew G Wilson

Hunter Elliott

Peyton Greenside