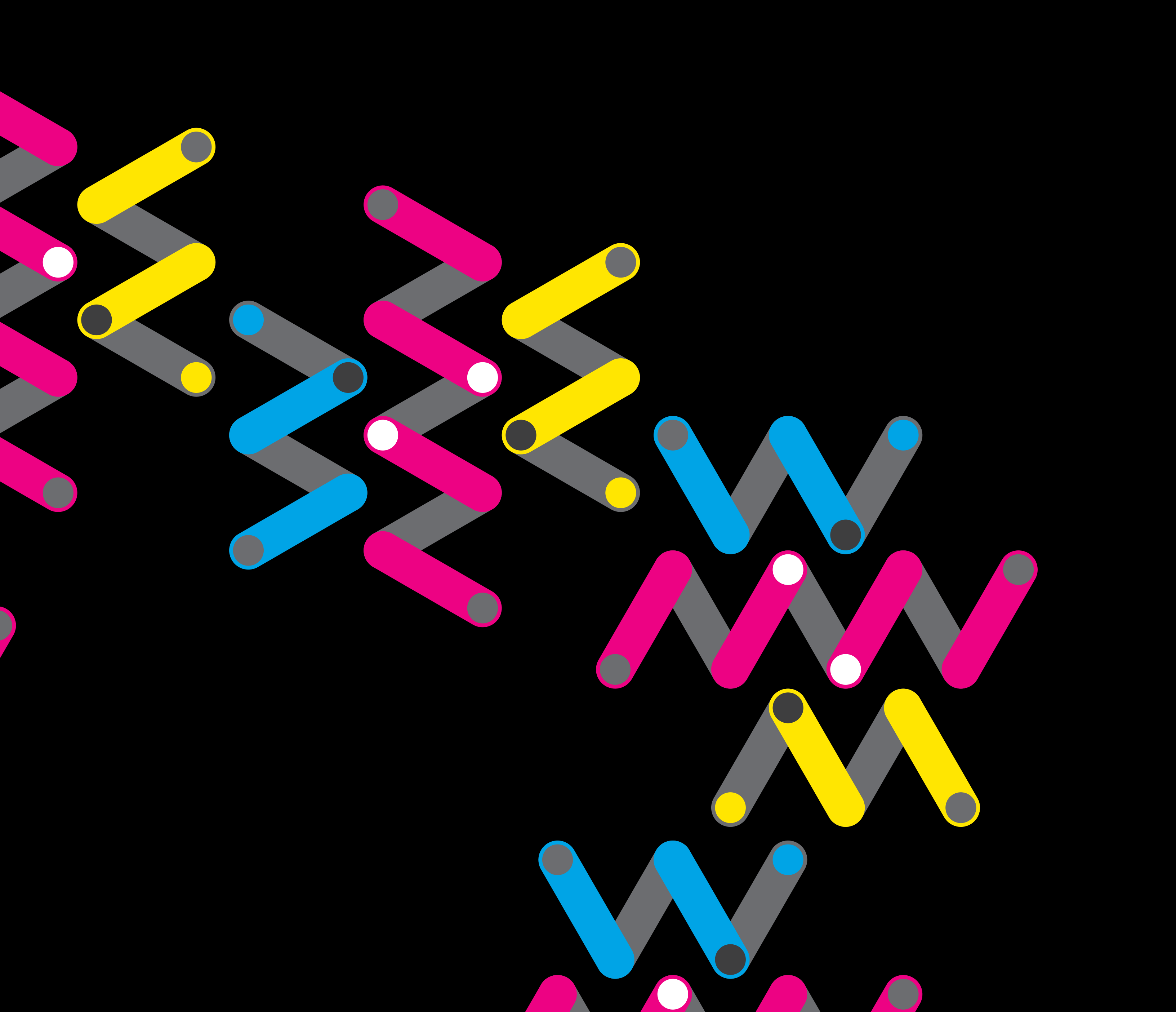


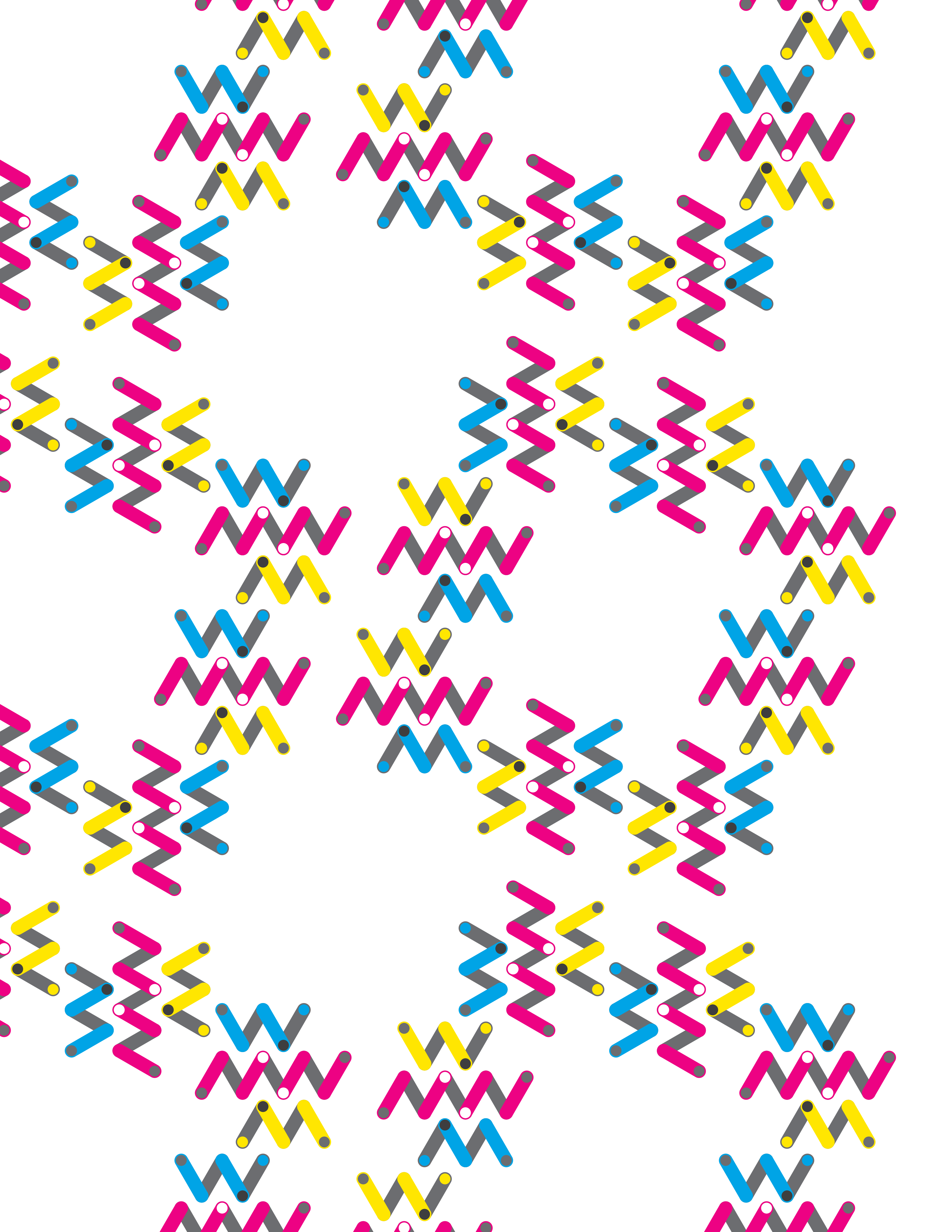
protein engineering

No surrogate assays, no compromises

Scaled for industrial success

Achieve your goals efficiently

ATUM



ATUM

protein engineering with proteinGPS

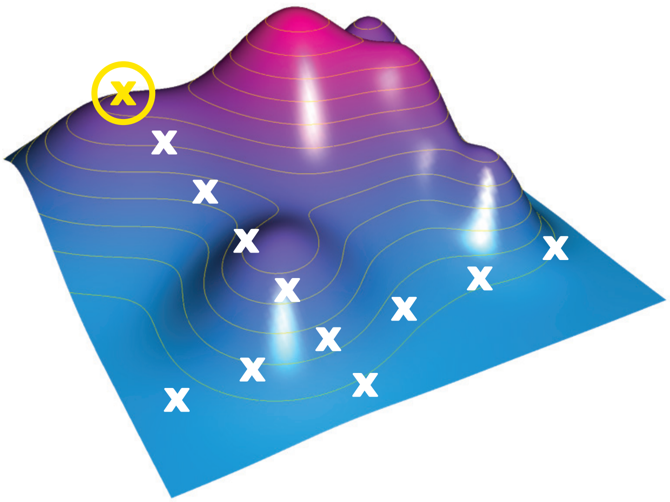
Protein properties such as activity, substrate specificity, expression yield, affinity, stability, aggregation, immunogenicity, and much more can be engineered through changes in the amino acid sequence of a protein or protein complex. Historically, there are two general strategies for protein engineering: rational protein design relying on a complete mechanistic understanding and directed evolution relying on more or less random search and selection.

ATUM has developed a unique protein engineering (ProteinGPS®) platform based on Machine Learning and Design of Experiment (DoE). The ProteinGPS proprietary technology uses similar megadimensional, empirical optimization algorithms as currently applied to gasoline formulation, web advertising, and stock market investing. ProteinGPS technology relies on DoE to calculate the set of nodes that are maximally information-rich in the relevant space, gene synthesis to make those exact sequences, high quality measurements to quantify function, and machine learning to find the optimal solution.

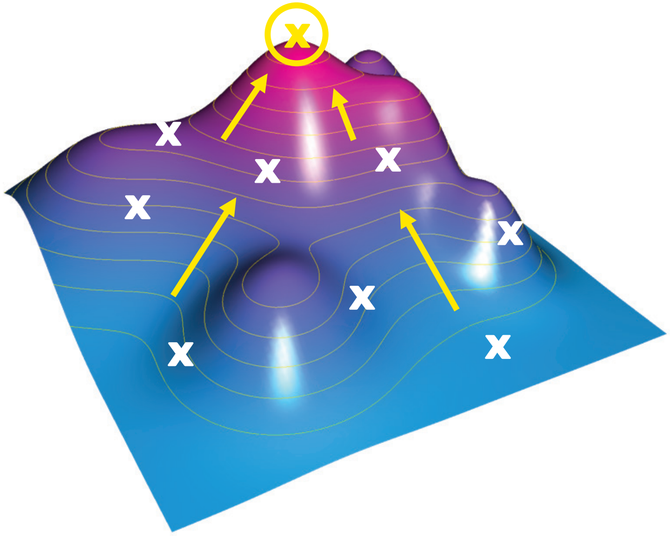
	Typical library approach	ProteinGPS engineering
Number to screen per round	10^4 - 10^{12} Limited by transformation frequency	48 - 96
Sampled space	<10^8	>10^{16}
Sampling of sequence space	Highly biased due to molecular biology process	Mathematically optimal
Assay requirement	High throughput, Typically requires a surrogate assay	Low throughput, high quality assay, Identical/similar to 'real' function
Multidimensional function optimization	Screening for only one function	Yes, Engineering for many functions in parallel
Redundant clones	Many	None
Learning algorithm	No, Usually pick best clone and repeat	Yes, Iterative expansion of comprehensive sequence-function map
Functional statistics	Fragile, substitutions not internally validated	Robust, substitutions validated in multiple systematic contexts
Engineering emphasis	Test	Design, Learn

design of experiment (DoE)

DoE relies on systematic variance to explore sequence-function correlation efficiently. The DoE process is faster, more efficient, captures multivariable interactions, and finds the best solution.

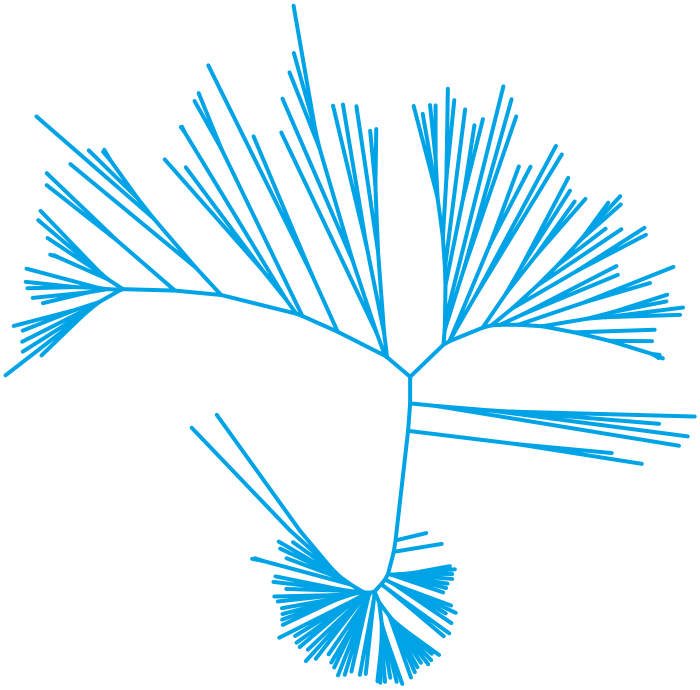


One Factor at a Time (OFAT) Optimization

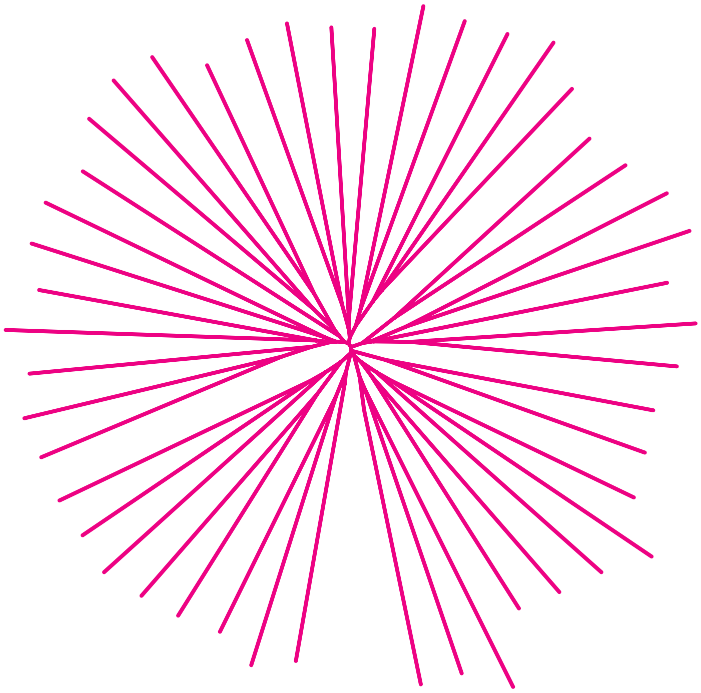


Design of Experiment (DoE) Optimization

Sequence diversity based on the pairwise Hamming distances in a typical random library (left) and a ProteinGPS dataset (right). The ProteinGPS gene variants (Infologs) enable efficient DoE based sampling of the entire sequence-function space.



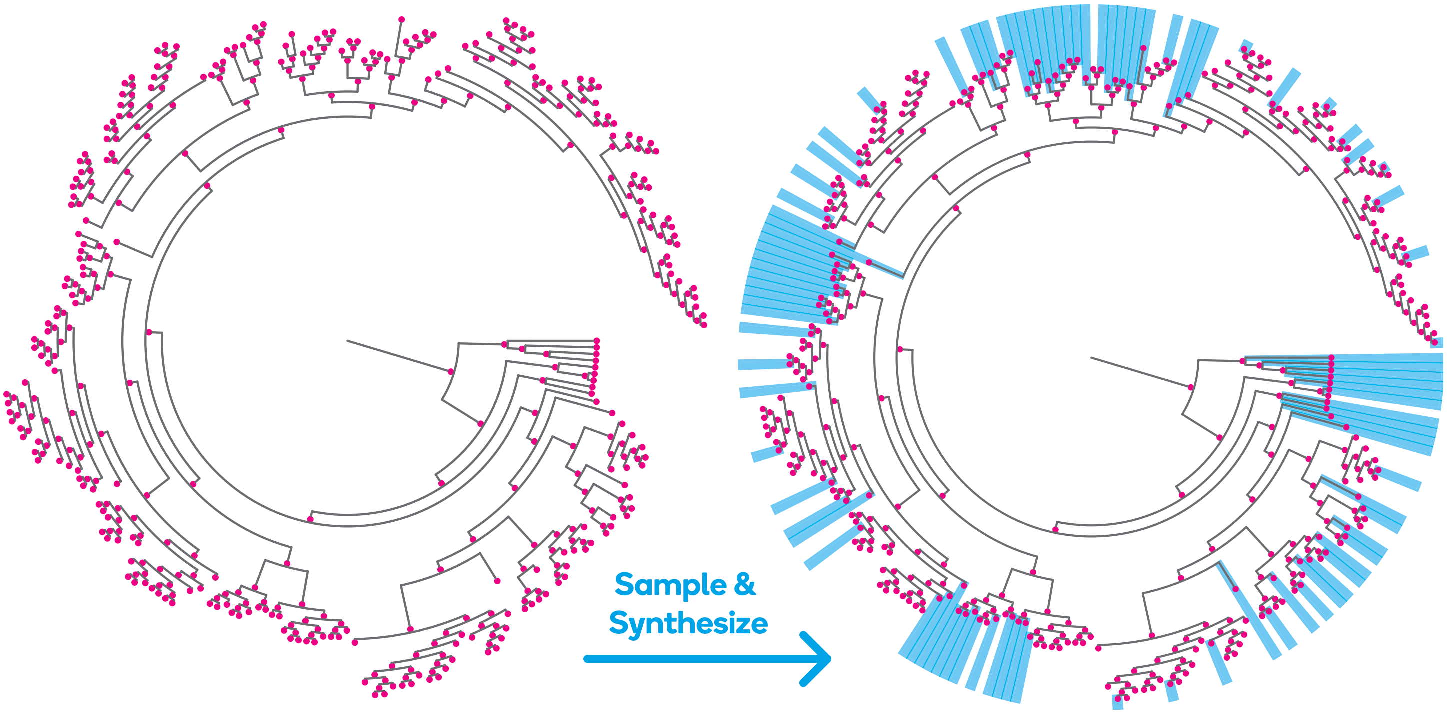
Random library



Infologs

finding the best starting point

Before initiating a ProteinGPS program, there is often a need to identify a good starting point. This step is particularly relevant for biocatalysis and similar applications, and less so for example protein therapeutics. ATUM has developed a standardized process to uniformly and accurately sample the phylogenetic tree of one or more protein families. The sampled sequence space is derived from public domain genetic databases and other sources.



Current and ancestral homologs are re-coded for host expression (e.g. E. coli) using ATUM proprietary technology and tagged for solubility and purification. The genes are synthesized, Electra-cloned, sequence verified, expressed, and the relevant functional activity(ies) assessed.

ATUM will develop an evolutionary parsed set of natural genes based on an understanding of molecular biology of the functional/sequence diversity. The homologs are explored for metagenomic distribution, multiple sequence alignment, phylogenetic trees and reconstructed ancestral sequences to correctly identify each sequence while avoiding data errors (sequencing errors, misalignments etc.

Depending on functional activity outcome from the homologs, it may be relevant to further drill down into one or more of the richer phylogenetic branches for synthesis and testing of additional related homologs. This second iteration is often useful for increased functionality and/or broader intellectual property claims.

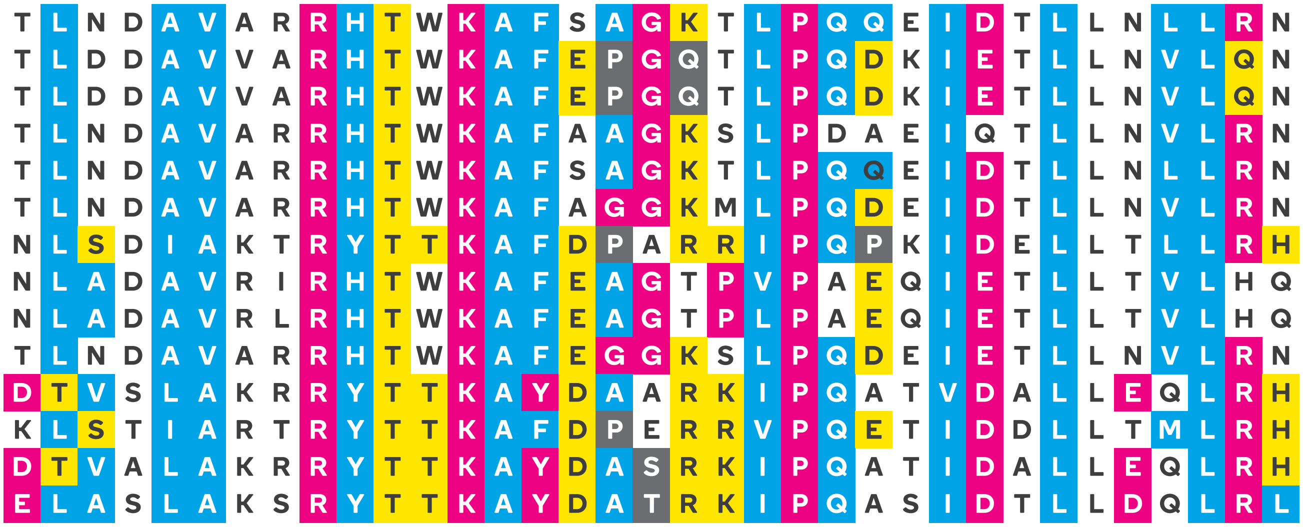
The ProteinGPS process may use one, two, or more starting points for the subsequent ProteinGPS engineering depending on the outcome of the phylogenetic search, the number and distribution of the functional properties to engineer, and any other constraints that could affect the search.

proteinGPS engineering process

All optimization can be divided into two key steps: Variable selection — how to choose amino acid substitutions to test, and Search — how to combine substitutions for best effect.

Variable selection

ATUM builds a complete alignment of all homologs of a given protein family centered on the starting point(s) and identifies all sequence diversity available. Each amino acid substitution in the alignment is assigned multiple scores based on evolutionary, structural, and functional analysis (if available). Scores for each substitution are averaged, normalized, and mean centered. Substitutions are rank-ordered accordingly and included for engineering.



```
T L N D A V A R R H T W K A F S A G K T L P Q Q E I D T L L N L L R N
T L D D A V V A R H T W K A F E P G Q T L P Q D K I E T L L N V L Q N
T L D D A V V A R H T W K A F E P G Q T L P Q D K I E T L L N V L Q N
T L N D A V A R R H T W K A F A A G K S L P D A E I Q T L L N V L R N
T L N D A V A R R H T W K A F S A G K T L P Q Q E I D T L L N L L R N
T L N D A V A R R H T W K A F A G G K M L P Q D E I D T L L N V L R N
N L S D I A K T R Y T T K A F D P A R R I P Q P K I D E L L T L L R H
N L A D A V R I R H T W K A F E A G T P V P A E Q I E T L L T V L H Q
N L A D A V R L R H T W K A F E A G T P L P A E Q I E T L L T V L H Q
T L N D A V A R R H T W K A F E G G K S L P Q D E I E T L L N V L R N
D T V S L A K R R Y T T K A Y D A A R K I P Q A T V D A L L E Q L R H
K L S T I A R T R Y T T K A F D P E R R V P Q E T I D D L L T M L R H
D T V A L A K R R Y T T K A Y D A S R K I P Q A T I D A L L E Q L R H
E L A S L A K S R Y T T K A Y D A T R K I P Q A S I D T L L D Q L R L
```

2

Design

DoE based mining of available sequence space and combining substitutions.

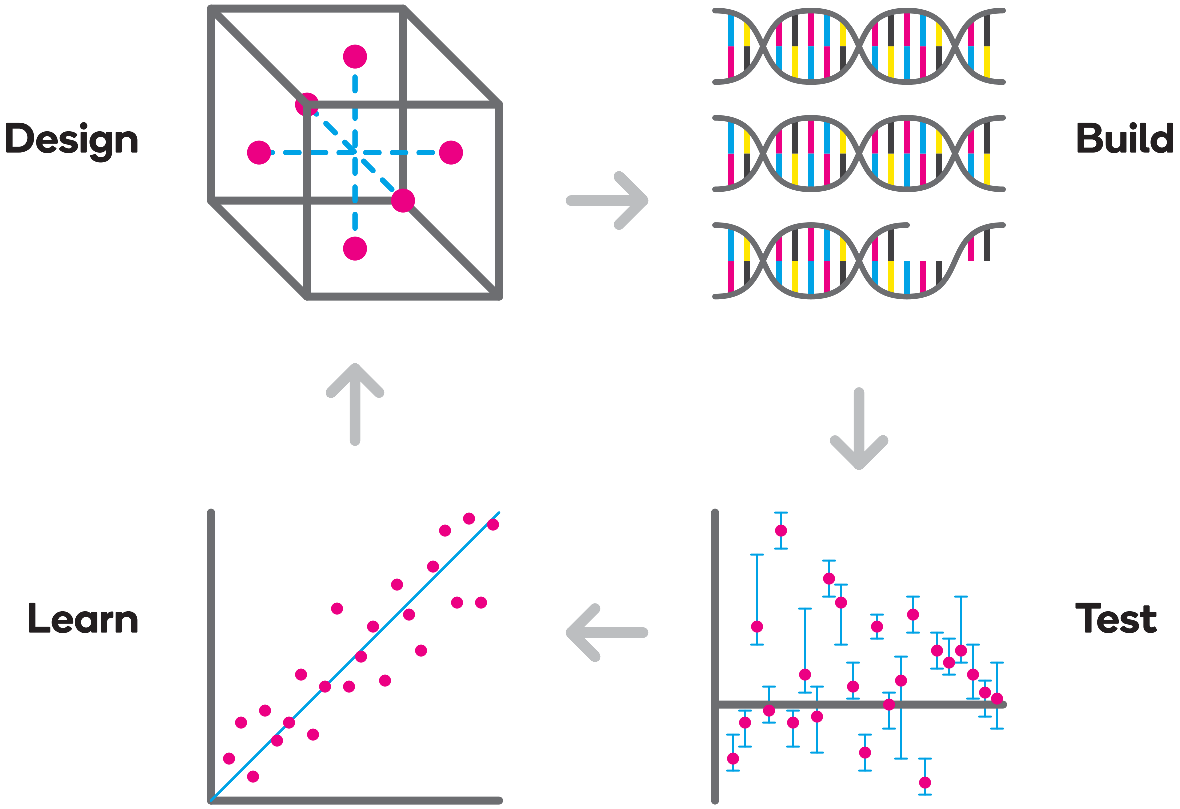
Build

Synthesizing individually designed Infologs (48-96 per round) ensures that the physical implementation is identical to the virtual design with no random mutations.

1

Search

The variables identified are incorporated in a systematically varied set of gene variants (a.k.a Infologs) centered around the starting point(s). In the first round each Infolog is typically 3-5 amino acid substitutions away from all other variants in the round, including the starting point. Each substitution is present in 4-6 Infologs to access additivity. The substitution distribution in the Infolog set is determined through DoE algorithms. This process allows for maximum search efficiency throughout the 'design-build-test-learn' cycle.



Test

Test in commercially relevant assays. Results allow us to deconvolute how substitutions within a protein sequence modify its function.

Learn

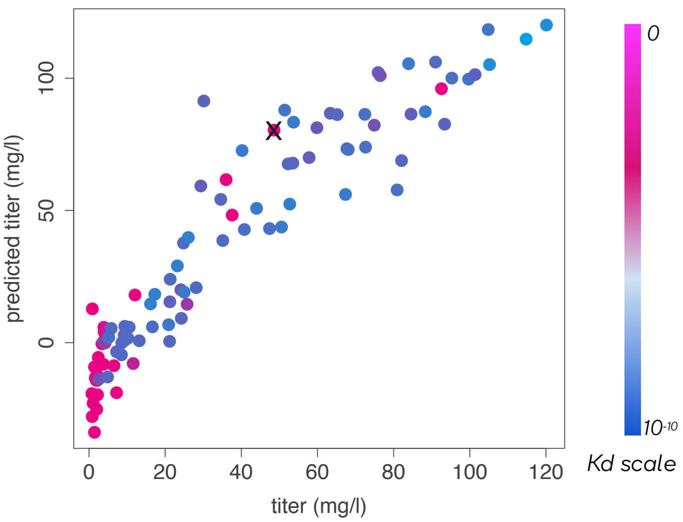
Establish a sequence-function model from the assay results and cross validate. Models are assessed based on their predictive value.

proteinGPS application examples

Engineering Antibodies for Developability

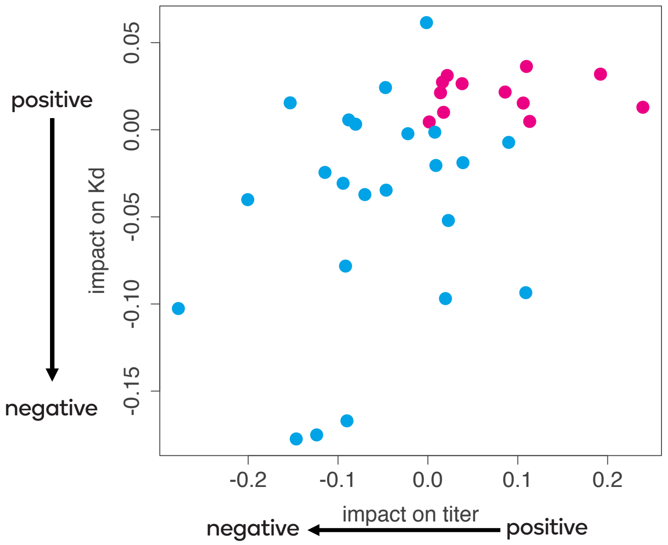
Phylogenetic and structural modeling identifies key residues affecting affinity, stability, expression yield, aggregation and humanization. The functional data derived from physical testing is modeled against the systematically varied infolog variants and used to generate predictions of new variants with enhanced developability properties.

Modeling titer



A total of 96 systematically designed variants of antibody X. The X-axis denotes the observed expression yield. The Y-axis denotes the predicted expression yield. The diagonal distribution represents the accuracy of the model. We have here color coded the binding affinity of the antibody variants. Parent (mouse) variant is denoted as 'X'.

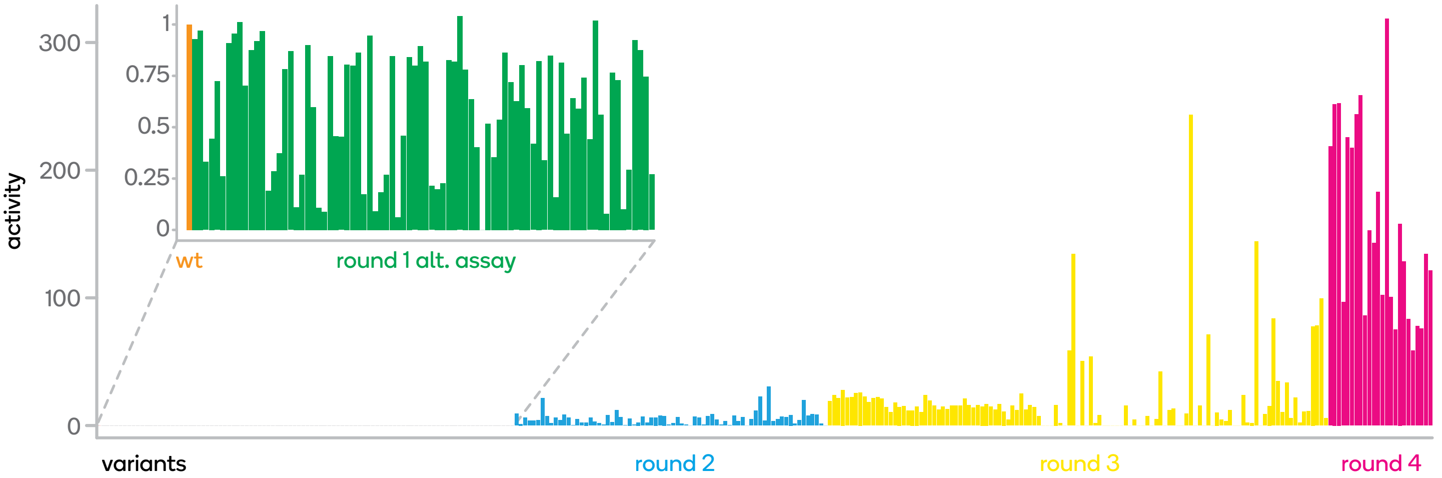
Impact of sequence variables



Amino acid substitutions used for humanization distributed by their relative contribution towards titer (X axis) and binding (Y axis). Substitutions contributing positively in both dimensions are denoted in pink.

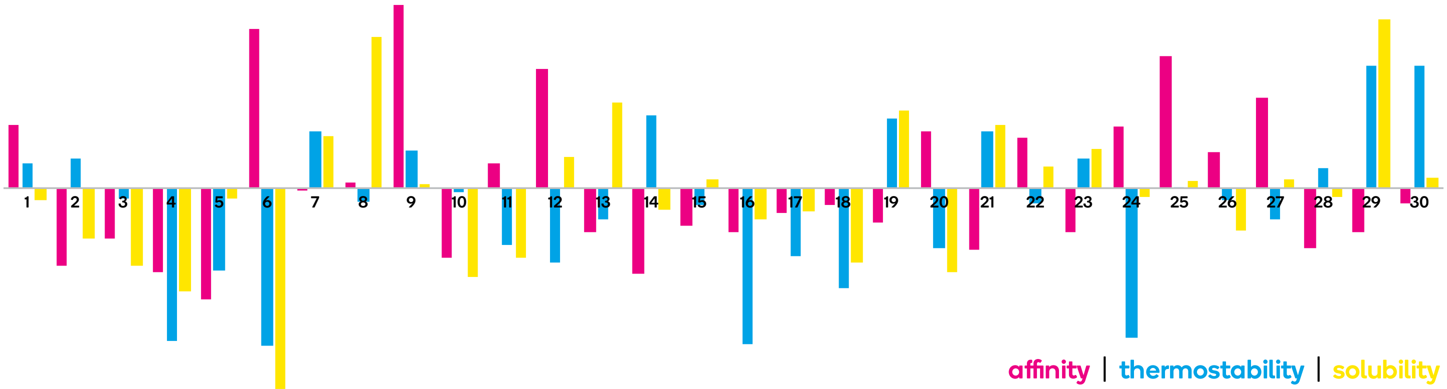
ATUM's proprietary Design of Experiment (DoE) technology enables systematic exploration of sequence-function relationships, identifying and quantifying amino acid substitutions and their relative contribution in multiple different functional dimensions. Assessing the sequence-function relationship and the amino acid substitutions relative independence provide guidance for generating predictive and testable models of target protein performance, metrics of its humanness, and developability. We typically test a total of 48-400 antibody variants over 1-4 iterations.

Process and Enzyme Engineering with Pfizer



Engineering of biocatalysis enzyme for Pfizer pharmaceutical intermediate synthesis. Four rounds (R1-R4) of biocatalytic variants screened for stereospecific activity for desired novel substrate. Several orders of magnitude improvement in specific activity was achieved testing a total of only 300 samples.

Multiple Functional Criteria



An industrial partner desired a protein with increased activity (blue bars), thermostability (red bars), and solubility (green bars) over their current candidate. ProteinGPS was used to characterize substitutions altering functionality in multiple dimensions. Combining positive effect substitutions ultimately produced several variants with orders of magnitude improvement in all 3 criteria.

References

ACS Synth Biol 2015. Mapping of amino acid substitutions conferring herbicide resistance in wheat glutathione transferase. Govindarajan et al.

Protein Eng Des Sel 2013. Redesigning and characterizing the substrate specificity and activity of Vibrio fluvialis aminotransferase for the synthesis of imagabalin. Midelfort et al.

PNAS 2010. Reconstructed evolutionary adaptive paths give polymerases accepting reversible terminators for sequencing and SNP detection. Chen et al.

J Biol Chem 2009. SCHEMA recombination of a fungal cellulase uncovers a single mutation that contributes markedly to stability. Heinzelman et al.

PNAS 2009. A family of thermostable fungal cellulases created by structure-guided recombination. Heinzelman et al.

Protein Eng Des Sel 2008. Protein engineering of improved prolyl endopeptidases for celiac sprue therapy. Ehren, et al.

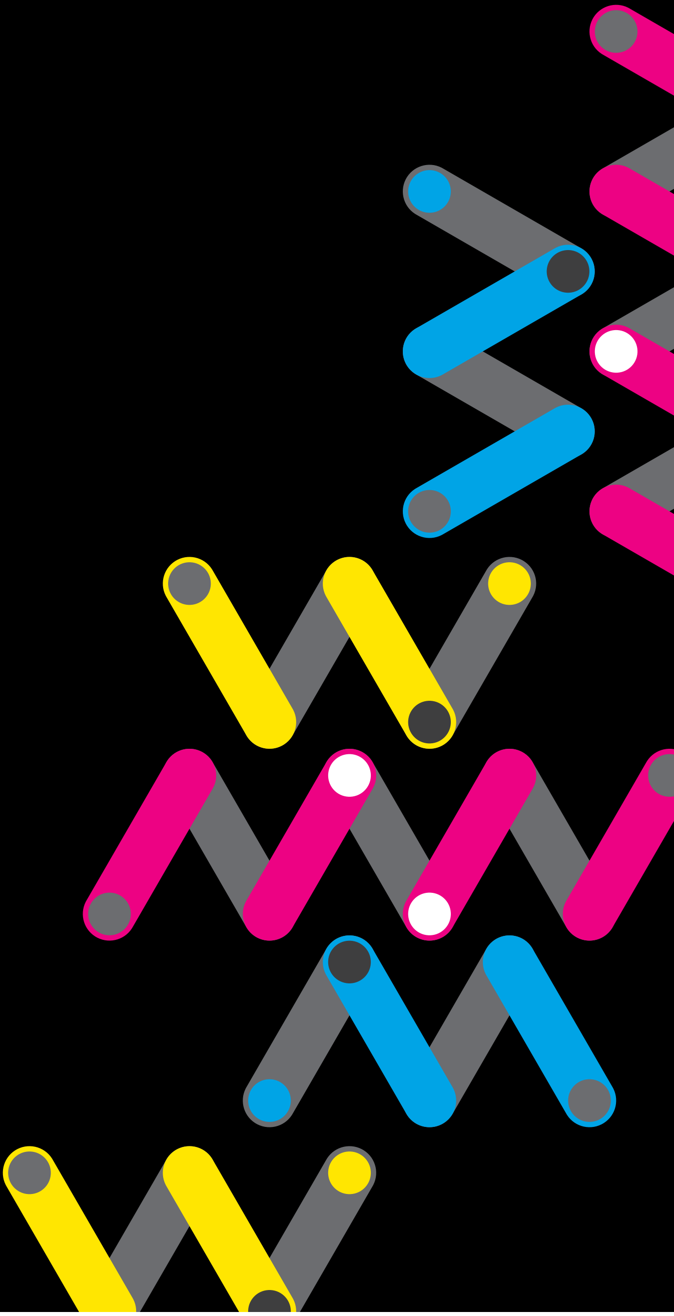
BMC Biotechnol 2007. Engineering proteinase K using machine learning and synthetic genes. Liao et al.

Curr Opin Biotechnol 2003. Putting engineering back into protein engineering: bioinformatic approaches to catalyst design. Gustafsson et al.

Patents

This technology is covered by issued US patents 8825411, 8635029, 8412461, 8401798, 8323930, 8126653, 8005620, 7805252, 7561973, 7561972, and related pending applications.

research. create. break through.



ATUM

+1 877 DNA TOGO
+1 650 853 8347
info@atum.bio