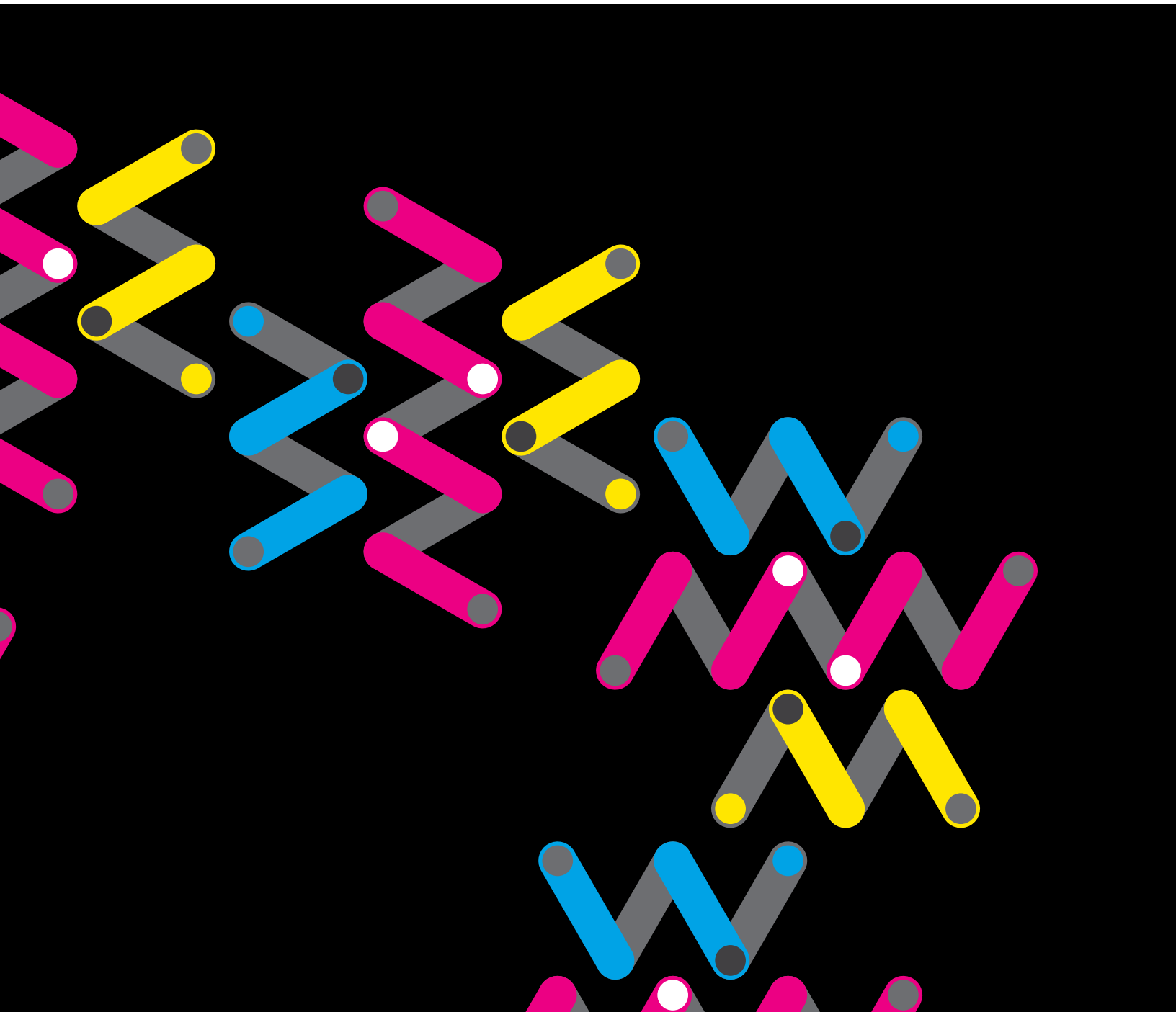
ATUM

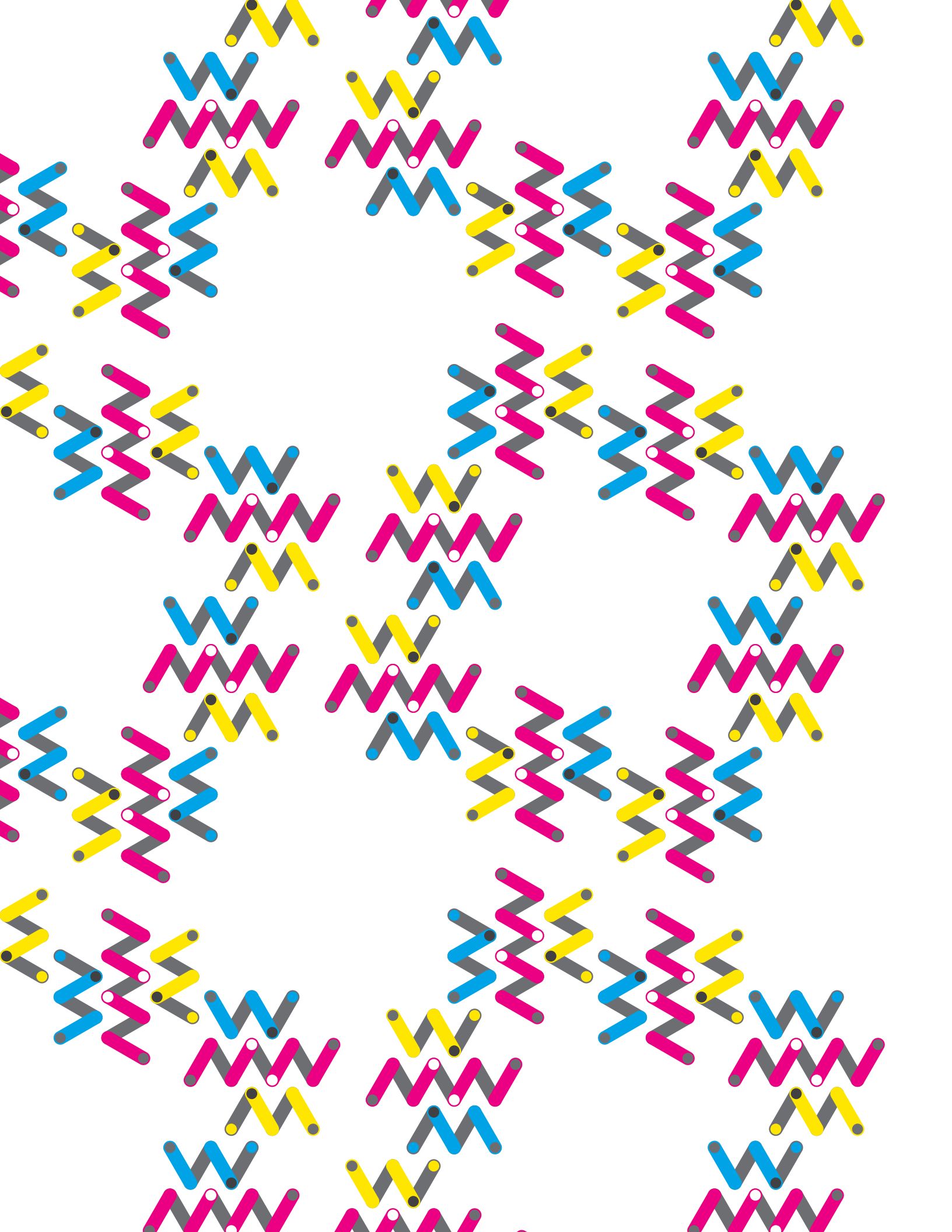
protein engineering

No surrogate assays, no compromises

Scaled for industrial success

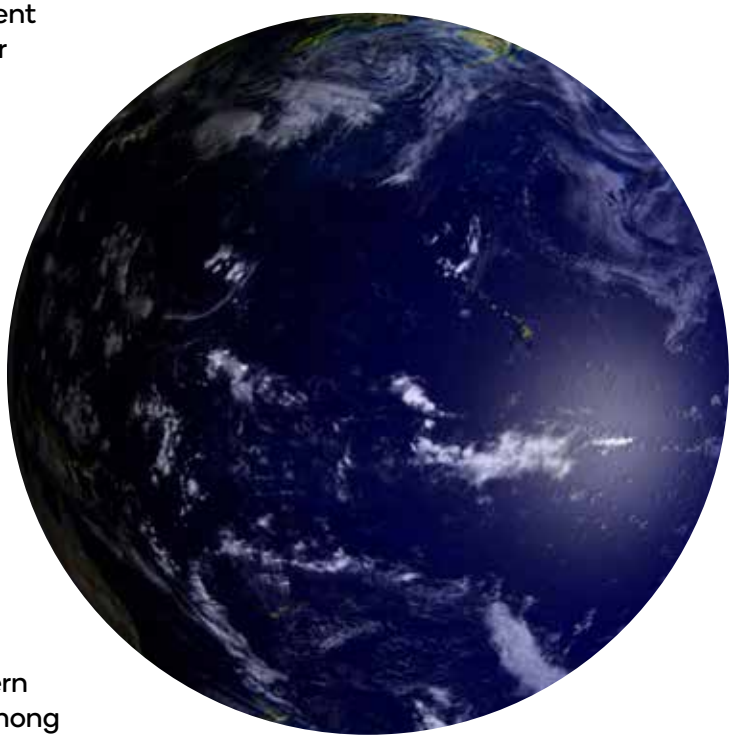
Achieve your goals efficiently





engineering proteins

Nature provides countless related but different proteins with largely untapped potential for commericial applications. Natural proteins are rarely optimal for their envisioned human use and require engineering to enhance performance. Commercial applications of the protein require simultaneous optimization in a number of correlated or orthologous functional properties.



The sequences of approximately 10^{12} different natural proteins can be found in various databases. Even though this is a large number, it is dwarfed by the number of all theoretically possible proteins 20^N (where N = number of residues in the protein). In a sequence-function ocean of possible sequences, there are only a very small number of activity islands that cluster around naturally existing proteins. Everywhere else, the sequences cannot even fold into stable structures, much less perform some function. ATUM uses pattern matching algorithms to ensure we only search among protein sequence islands that support activity. We let phylogenetic information guide our ProteinGPS® algorithms to efficiently explore islands of biological life to engineer proteins.

Protein properties such as activity, substrate specificity, expression yield, affinity, stability, aggregation, immunogenicity, and much more can be engineered into natural protein sequences through changes in the amino acid sequence of the protein or protein complex.

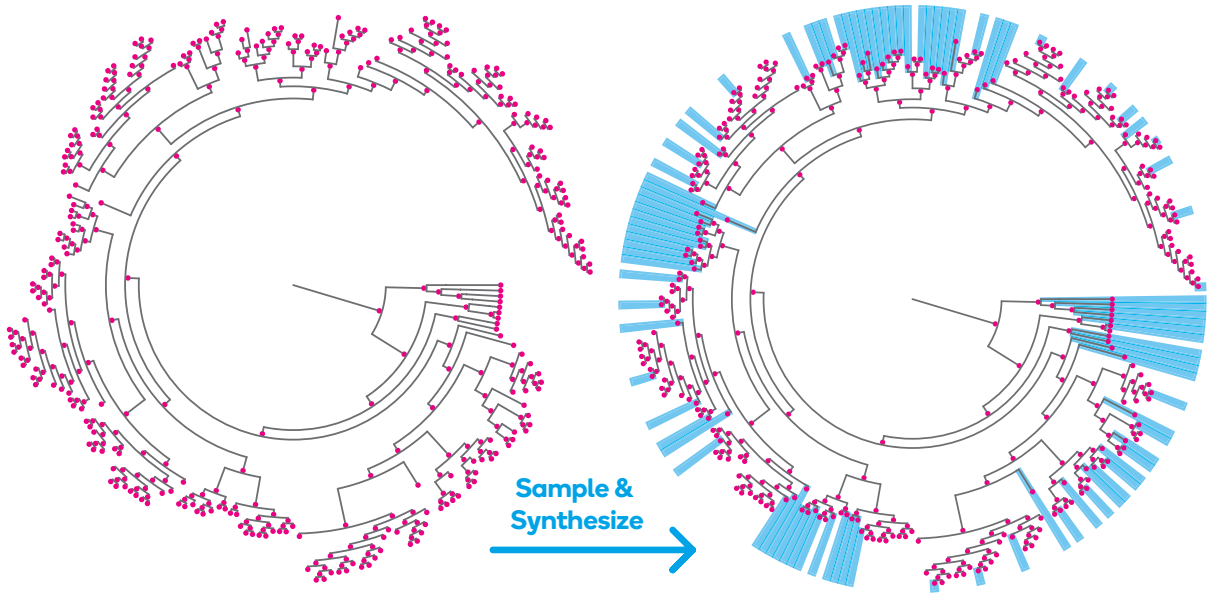
ATUM's unique protein engineering platform ProteinGPS® is based on Machine Learning and Design of Experiment (DoE). The ProteinGPS proprietary technology uses the same megadimensional, empirical optimization algorithms currently applied to diverse aplications such as gasoline formulation, web advertising, and stock market investing. The 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 sequence nodes, high quality measurements to quantify function, and machine learning to find the optimal solution.

"Tapping into the vast diversity of proteins found in nature combined with machine learning is a key aspect of ATUM's unique protein engineering platform ProteinGPS® to offer our customers the opportunity to obtain proteins with optimal properties for commercial applications."

Dr. Claes Gustafsson, Co-Founder & CCO

bioprospecting - 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 protein therapeutics. ATUM has developed a 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.

For every new project, 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 correcting for data errors such as 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.

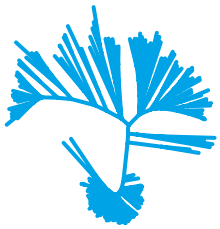
Natural selection guides our ProteinGPS choice of amino acid substitutions. We score and rank every possible amino acid substitution using several different evolutionary, structural and functional parameters. Finally, we apply DoE to design protein variants incorporating the highest ranked substitutions.

why GPS is better than a library approach

Design

Library

10^4 - 10^9 variants

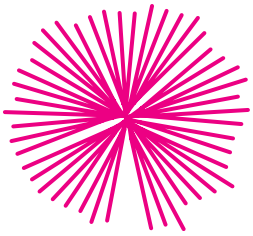


Many more random gene variants need to be used to infer sequence-function correlation. Impossible to separate correlation from causality.

Imbalanced and incomplete sampling

GPS

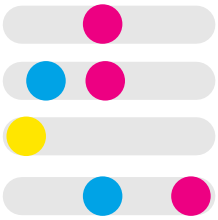
96 variants



Fewer systematically varied gene variants cover much more sequence-function space. Direct causality is captured. Epistatic effects are quantified.

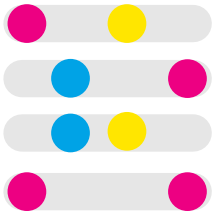
Balanced sampling of sequence space

Build



Random

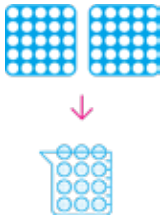
No control over which mutations occur together. There is thus no means by which to understand any interactions between mutations. Random mutations, deletions and nonsense mutations are sprinkled throughout, further confounding interpretation.



Systematic

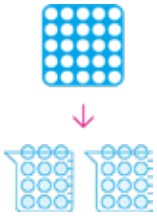
Using DoE, each substitution is represented multiple times. This allows us to subsequently determine whether any substitutions do not play well with others.

Test



Large libraries require the use of high throughput surrogate screens. These typically measure only one property, often under conditions that are very different from the intended application.

High throughput screens, only one property measured



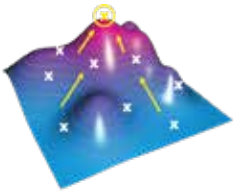
12-96 Infologs are built, thereby enabling small, precise measurements in conditions representative of real-world applications.

Few assays measuring multiple properties

Learn

Classic inefficient search

Library screening typically starts by measuring one property and selecting all library members whose activity exceeds a certain threshold. Those survivors are then screened for a second property and so on. There is no learning, so no optimization of combinations is possible.



Efficient search

Different sequence changes frequently affect different properties, and often a change that is good for one desired activity is bad for another. By quantifying the effects of every sequence change on every property, it is possible to select combinations of changes that alter all properties in the desired direction.

proteinGPS® engineering process

Optimization can be divided into two key steps:
A. Variable selection - which amino acid substitutions to test, and
B. Search - how to combine substitutions for best effect.

A. Variable selection

Natural selection guides our ProteinGPS® choice of amino acid substitutions or variables.

To identify variables, ATUM builds a complete alignment of all homologs (often many thousands) of a given protein family centered on the starting point(s) and identifies all sequence diversity available within that sequence island. A typical 1 kb gene may have ~1000 optional substitutions. 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 and the top 50-100 substitutions are included for engineering.

2

Design

Computational 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.

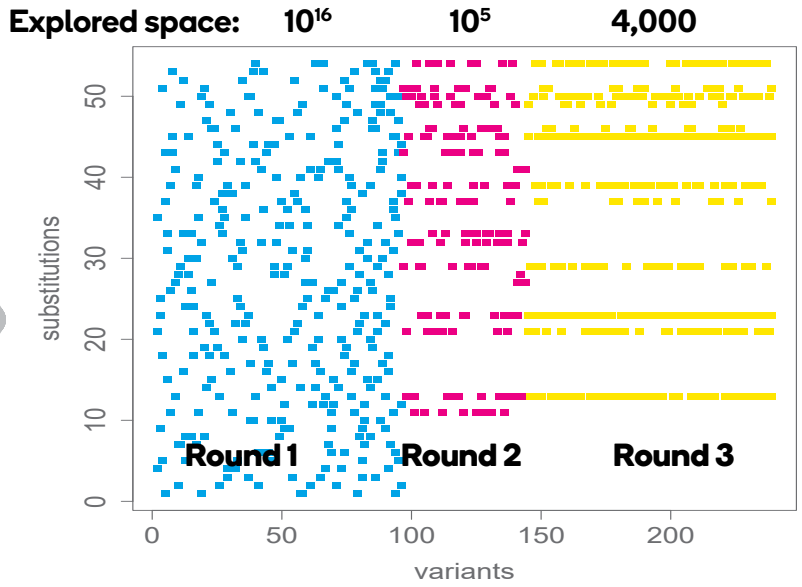
1

B. Search

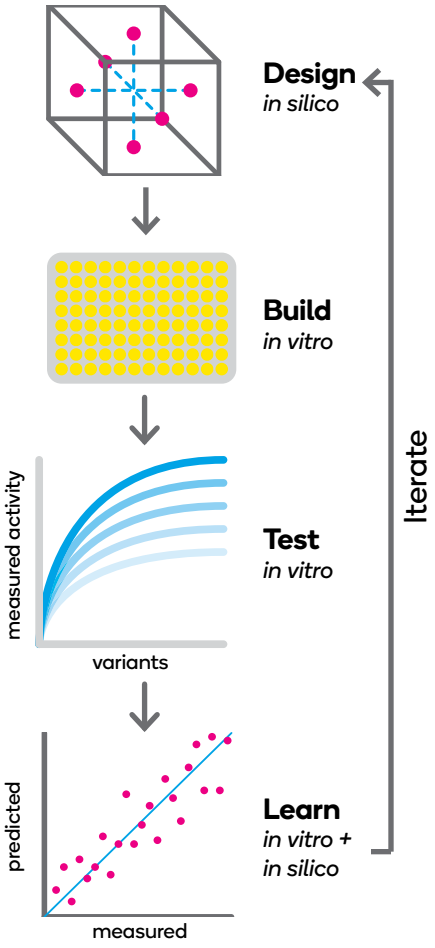
The variables (amino acid substitutions) identified are incorporated in a systematically varied set of gene variants (a.k.a Infologs) centered around the starting point(s). In the first round each Infolog is typically 3-5 amino acid substitutions away from all other variants in the round, and each substitution is present in at least 10-30% of the Infologs. 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.
Fewer than 100 protein variants are typically needed for each ProteinGPS iteration.
Once the protein properties for the intended application have been measured, ProteinGPS machine learning is then used to determine the effect of each amino acid substitution independently and in context.

ProteinGPS® allows us to identify and combine changes that separately improve desired properties, resulting in a protein that is improved for all desired properties.



Each round incorporates best variants from the previous rounds and builds on these to make a new set through DoE algorithms.



4

Test

Test in commercially relevant assays.

Learn

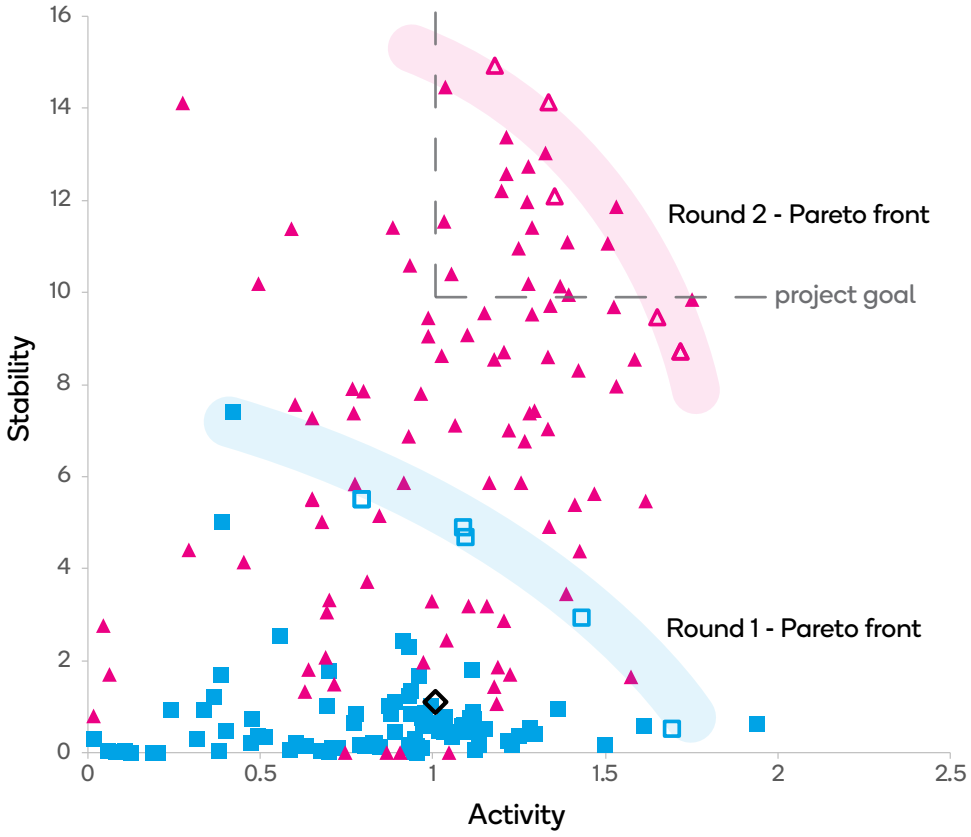
Establish a sequence-function model from the assay results and use cross-validation to assess predictibility.

3

case studies: proteinGPS®

Engineering Proteins for Activity and Stability

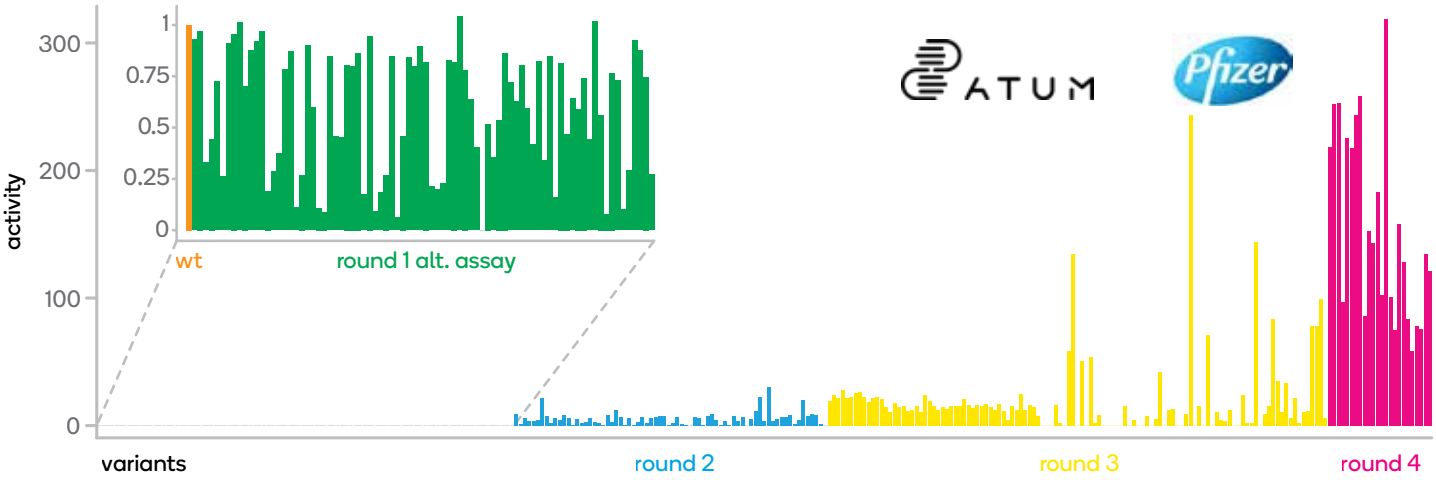
Phylogenetic, functional and structural modeling identifies key residues affecting affinity, stability, expression yield, aggregation and any other relevant property. The functional data is derived from physical testing and is modeled against the systematically varied Infolog variants. The models are used to generate predictions of new variants with enhanced properties.



Commercial goals of stability and activity shown by dotted line were reached by ~20 different variants after making and testing a total of only 190 systematically varied proteins, 95 in each round. Graph shows activity (*x-axis*) and stability (*y-axis)* for rounds 1 (blue squares) and 2 (pink triangles), data has been normalized to wild type. The blue outline squares represent variants predicted in our models to represent best possible output from 30,000 combinations of 3 substitutions in round 1 i.e. round 1 'pareto front' (mathematical evaluation of best output(s) across multi-dimensional outputs). The pink outline triangles represent variants predicted to represent best possible output from 30,000 combinations of 3 additional substitutions in round 2 i.e. round 2 pareto front. Black outline diamond denotes parent.

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, stability, and developability. We typically test a total of 48-400 protein variants over 2-4 iterations.

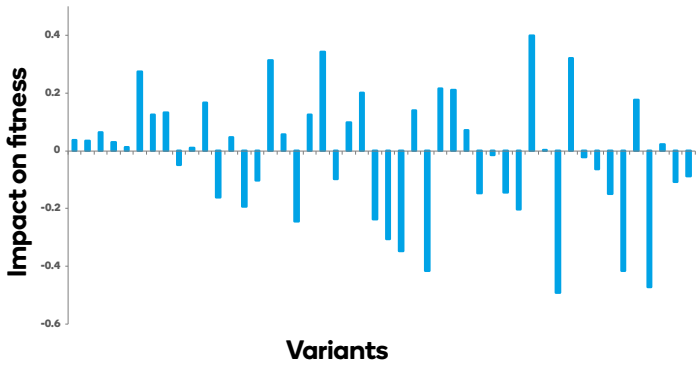
Process and Enzyme Engineering with Pfizer

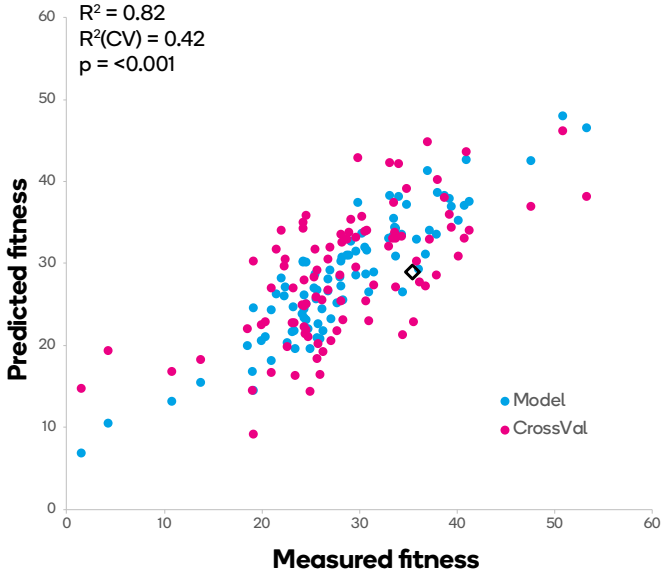


Engineering of biocatalysis enzyme for Pfizer pharmaceutical intermediate synthesis. Four rounds (R1-R4) of biocatalytic variants screened for stereospecific activity for desired novel substrate. Several orders of magnitude improvement in specific activity was achieved while testing a total of only 300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

Life is multidimensional

Σ Fitness =	6x Function A	2x Function E
	1x Function B	2x Function F
	3x Function C	5x Function G
	5x Function D	5x Function H





An industrial partner requested an enzyme improved on 8 separate functional dimensions. The relative commercial value of each dimension varied over six fold as shown in table. ATUM built and executed on an eight dimensional integrated fitness function (graph on left) with an overall cross-validation R^2 *>0.4*. Combining positive effect substitutions ultimately produced several variants with orders of magnitude improvement in all criteria. Graph on right shows good correlation between predicted and measured fitness with orders of magnitude improvement over parent (black outline diamond) in all 8 functional criteria.

"Our bioengineering ProteinGPS® technology has now been validated through many scientific collaborations, peer-reviewed scientific publications and strong underlying IP. It is exciting to see broad adoption of the technology with partners ranging from Fortune 500 companies to small startups, and in commercial fields ranging from chemical biocatalysis to therapeutic proteins"

Dr. Sridhar Govindarajan, Co-Founder and CIO

	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
Epistatic effects	Ignored	Identified and quantitated
Assay requirement	High throughput, Typically requires a surrogate assay	Low throughput, high quality assay, Identical/similar to 'real' function
Multidimensional function optimization	No, Screening for only one function	Yes, Engineering for many functions in parallel
Redundant clones	Many	None
Non-functional clones	Many	Few
Learning algorithm	No, 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	HTP screen	Design, Learn

How do customers benefit from ProteinGPS®

- Process optimization - find a better protein/enzyme solution to the current one
- Freedom to operate - identify proteins outside of existing IP restrictions
- Decrease risk by optimizing all relevant properties simultaneously
- Real world performance - product that performs better in the real world - from manufacturing to its end-application
- Speed to market - efficient workflow accelerates project timelines and positions your protein for fast scale-up and manufacturing

"ATUM's ProteinGPS® technology helped us quickly identify substitutions altering functionality in multiple dimensions, ultimately producing protein variants with orders of magnitude improvement in all 3 desired criteria and in the shortest possible timeframe."

Customer at rapidly growing protein pharma biotech company

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.

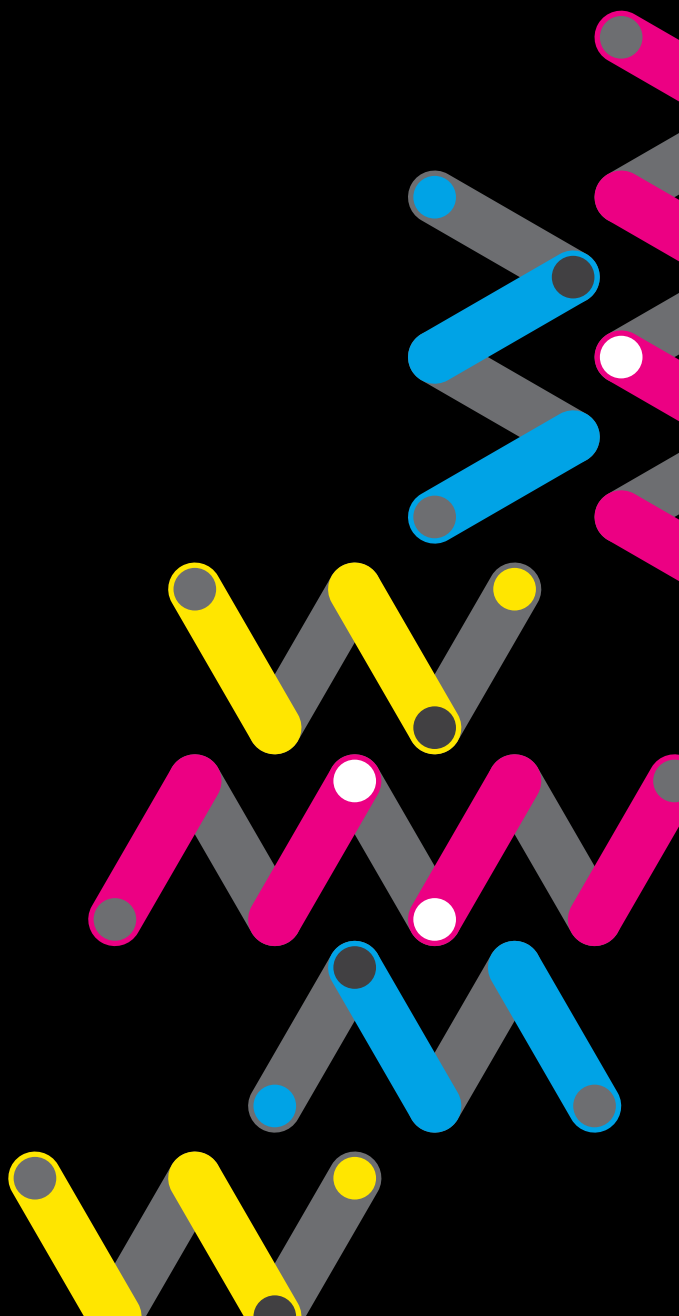
Dupont. US Pat. Nos. 9388392 and 9169467. Ketol-acid reductoisomerase enzymes and methods of use. Govindarajan et al.

ADM. US App. No. 20180112244. A genus of epimerase enzymes for conversion of fructose to allulose at high temperature and low pH. Venkitasubramanian et. al.

ATUM Patents

The technology described in this document is covered by issued US patents 10253321, 9206433, 9102944, 8825411, 8635029, 8412461, 8401798, 8323930, 8126653, 8005620, 7805252, 7561973, 7561972, and related pending applications.

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



research. create. break through.