BIOTRANS
LA ROCHELLE 2023

N°1503 / PC
TOPIC(s) : Enzyme discovery and engineering / Artifical intelligence / computational methods

Basecamp Research - Predictive Enzyme Development by Nature-Inspired AI

AUTHORS

Ahir PUSHPANATH / BASECAMP RESEARCH, UNIT 510 CLERKENWELL WORKSHOPS, 27 CLERKENWELL CLOSE, LONDON
Molly STEADMAN / BASECAMP RESEARCH, UNIT 510 CLERKENWELL WORKSHOPS, 27 CLERKENWELL CLOSE, LONDON
Valerio PERENO / BASECAMP RESEARCH, UNIT 510 CLERKENWELL WORKSHOPS, 27 CLERKENWELL CLOSE, LONDON

PURPOSE OF THE ABSTRACT

Biocatalysis is approaching an inflection point, yet enzyme discovery and evolution still remain bottlenecks for its accelerated adoption across all industries. Directed evolution, the hallmark technique for enhancing natural enzymes to meet industrial process specifications, has provided multiple success stories, but is inherently a resource intensive endeavor. Consequently, there is an increasing need for computational tools to reduce enzyme development times. As the intersection of AI and biology deepens, AI protein design tools are in the spotlight. Despite great advances built on the success of large language models, the protein sequence datasets used for training these algorithms are not diverse or large enough. Therefore, in order to usher in the desired paradigm shift of truly predictive design, a true representation of the protein landscape is required.

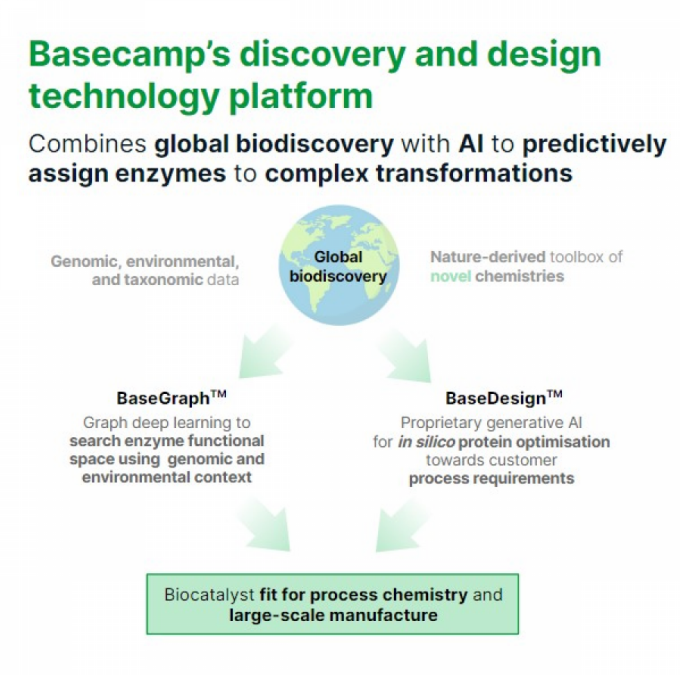
Basecamp Research's global biodiscovery has yielded hundreds of millions of ethically sourced, novel protein sequences from diverse eco-regions, resulting in a dataset three times larger and four times more diverse than Uniprot. For commonly used biocatalyst classes such as IREDs and KREDS, BaseData™ has over 25 times more than those available on UniProt, many sourced from extreme environments. For every sample collected, extensive environmental metadata is also recorded. Our unique data on nature is not stored in a tabular format but rather as a pre-indexed, interconnected network, incorporating over 3 billion relationships between proteins, genomes, taxonomic communities, and environmental metadata. Using graph deep learning techniques, our BaseGraph™ algorithm can intelligently navigate the protein landscape, using the evolutionary context of a protein to infer complex function and process performance with unparalleled accuracy.

Our BaseDesign™ algorithms herald a new vision for enzyme optimisation. By fine tuning protein language models on BaseData, our generated sequences have lower perplexity, higher pLDDT and TM scores compared to other protein design tools trained exclusively on public databases1.

The success of our methods are exemplified in two customer case studies. In the first study, Basecamp Research identified a novel broad specificity wildtype transaminase within 1 week using BaseGraph™. The same broad specificity was only achieved after 1.5 years and > £1 million in R&D costs with traditional directed evolution . In the second study, BaseDesign™ was used to generate novel fluorinase enzyme sequences, a rare enzyme class with only 18 representatives in the public database. Some of the generated sequences had 90 mutations over two wildtype literature controls. Out of the 66 novel sequences tested, remarkably, 93% of them were expressed for the first time in the expression host of choice, with 82% of them exhibiting fluorinase activity. Additionally, 64% of the designed sequences outperformed the two literature control sequences in SAM fluorination activity and

thermostability.

FIGURES



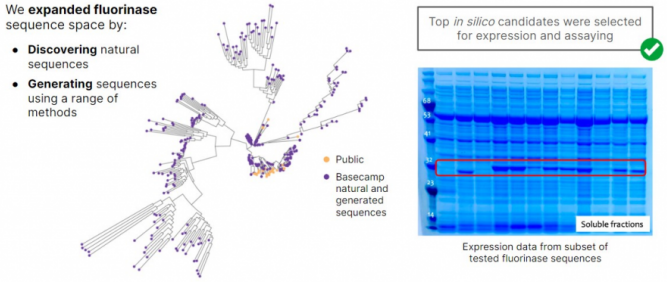


FIGURE 1

Overview of Technology

BaseData: Unveiling the True Protein Universe.

BaseGraph: A Search Engine for Proteins, using Evolutionary Context to Infer Complex Function and Performance.

BaseDesign: Best in Class Protein Design AI.

FIGURE 2

Combination of nature and AI expands the limited sequence pool of a rare biochemistry, fluorination.

Expanding Fluorinase Sequence Space with Nature and AI: Basecamp Research Generates Novel Sequences with High Expressibility, Thermostability, and Activity for SAM-Fluorination.

KEYWORDS

Generative AI | Protein Design | Metagenomics | Transaminase and Fluorinase

BIBLIOGRAPHY

[1] Munsamy, Geraldene, et al. "ZymCTRL: a conditional language model for the controllable generation of artificial enzymes."