BASECAMP RESEARCH : PREDICTIVE ENZYME DEVELOPMENT THROUGH NATURE AND AI

Ahir Pushpanath, Basecamp Research
ahir@basecamp-research.com
Valerio Pereno, Basecamp Research
valerio@basecamp-research.com

Key Words: Generative AI, Global Metagenomics, Graph Deep Learning ,Transaminase, Fluorinase

Biocatalysis is at a turning point, but enzyme discovery and evolution continue to pose significant challenges. Directed evolution, the primary method for enhancing natural enzymes to meet industrial process specifications, is a resource-intensive process. Consequently, computational tools are urgently needed to reduce enzyme development times. As AI and biology intersect, AI protein design tools are gaining more attention. However, despite progress built on large language models' success, protein sequence datasets used for training these algorithms are not diverse or large enough. To usher in the desired paradigm shift of truly predictive design, a true representation of the protein landscape is necessary.

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. BaseData™ has over 25 times more sequences in commonly used biocatalyst classes, such as IREDs and KREDs than UniProt, many of which are sourced from extreme environments. For every sample collected, extensive environmental metadata is recorded. Our unique data on nature is stored in a pre-indexed, interconnected network, encompassing over 3 billion relationships between proteins, genomes, taxonomic communities, and environmental metadata. Using graph deep learning techniques, our BaseGraph™ algorithm intelligently navigates the protein landscape, inferring complex function and process performance with unparalleled accuracy based on the evolutionary context of a protein.

Our BaseDesign™ algorithms offer a new approach to enzyme optimization. By fine-tuning protein language models on BaseData™, we generate sequences that exhibit lower perplexity, higher pLDDT and TM scores compared to other protein design tools trained exclusively on public databases.
The success of our methods is demonstrated in two customer case studies. In one 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% were expressed for the first time in the expression host of choice, with 82% exhibiting fluorinase activity. Additionally, 64% of the designed sequences outperformed the two literature control sequences in SAM fluorination activity and thermostability.

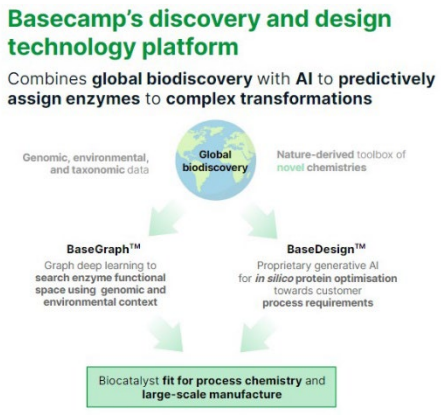

Figure 1 : BaseData™: Unveiling the Protein Universe, BaseGraph™: Evolutionary Context Search for Proteins and BaseDesign™: Precision Protein Design AI.