

Machine learning applied to biocatalysis research



Machine learning (ML) is rapidly influencing the development of many research areas, including biocatalysis, the use of enzymes and living systems to mediate chemical reactions, often of pharmaceutical and industrial importance. To find out how can machine learning be applied to biocatalysis research, we talked to three scientists, Professor Rebecca Buller (Zurich University of Applied Sciences), Dr. Stanislav Mazurenko (Masaryk University) and Associate Professor Yang Yang (University of California, Santa Barbara). We discussed the potentials and challenges associated with ML, how collaborations between computational and biocatalysis experts can be promoted, and how will the field develop in the future.

What are your main research interests and what got you interested in applying machine learning to biocatalysis?

Rebecca Buller: Early on in my career I became intrigued by the potential of enzymes and the possibility to engineer Nature's catalysts for applications in medicine, environmental sciences and chemical manufacturing. My current research focusses on creating enzymes for challenging chemical



transformations, such as asymmetric halogenations, and understanding the factors that determine the biocatalysts' stability, selectivity and activity. In addition, my team develops efficient methods to engineer enzymes. For example, we used stability predictions to exclude deleterious mutations from enzyme library design accelerating the evolution of a de novo designed Kemp eliminase¹ or applied machine learning (ML) to guide the optimization of several enzymes, including a halogenase² and a ketoreductase³. ML enables us to explore large datasets and to analyze the sequence-function relationship of screened enzyme variants. In this way, we can navigate the protein fitness landscape more effectively.

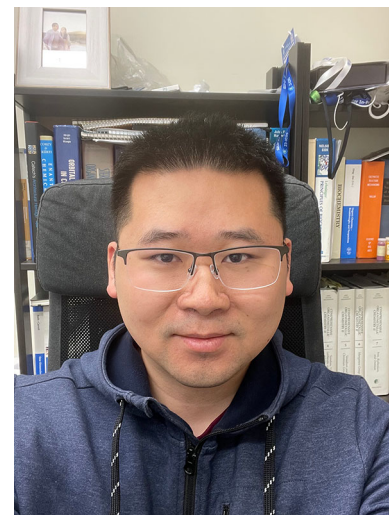
Stanislav Mazurenko: I come from a mathematical background focused on data analysis and modeling. In my early projects in biocatalysis, I worked on enzyme kinetics and thermodynamics—subfields that rely heavily on data modeling, dynamic equations, and parameter estimation to reveal enzymatic mechanisms from experimental data. The robust model selection there poses significant challenges due to many factors such as experimental noise, the presence of unknown or unobservable variables, numerous potential underlying mechanisms, etc. So, when machine learning (ML) began gaining traction in biology, I immediately got curious to try it out for our data.

The concept of learning a model directly from the data was—and still is—fascinating to me, especially for biomolecular systems, which are often too complex and vast for traditional, first-principle-based models to adequately capture. The progress in the domain has been breathtaking over the last five years. Every time a novel machine learning model is published that pushes the boundaries of what's possible in biocatalysis, I'm reminded of the immense potential of this technology. We now use ML in a wide range of applications: generating novel biocatalysts with protein language models, predicting the effects of mutations on various properties, such as stability and solubility, analyzing molecular dynamics simulations with artificial neural



networks, automating experimental data collection and analysis etc⁴. We are also investigating how to design protein dynamics with ML and how to leverage emerging quantum computing methods in our pipelines.

Yang Yang: Our interest spans the broadly defined field of biocatalysis, with an emphasis on the design, discovery and development of enzymatic activities not known in nature, and recently, enzymatic activities not known in either organic chemistry or biochemistry. In one area of research, by exploiting the innate redox properties of first-row transition metals, we are repurposing and evolving metalloenzymes to catalyze stereoselective radical reactions not known in nature. In the other area of research, we are developing pyridoxal radical biocatalysis to advance stereoselective intermolecular free radical reactions which are not known in either chemistry or biology. We are also interested in



developing generally applicable machine learning guided enzyme engineering to speed up the optimization of useful biocatalysts.

How can machine learning aid the discovery, optimization, and engineering of biocatalysts?

Rebecca Buller: Biocatalysis is an approach to synthetic chemistry in which enzymes carry out chemical reactions. The process of creating an efficient biocatalyst involves the identification of a suitable enzyme with some level of starting activity for the desired transformation and optimization of the starting scaffold, usually via directed evolution, to tailor enzymatic properties.

In the recent past, the number of available protein sequences has increased by a staggering 20-fold (2023: > 2.4 billion^{5,6}; 2018: ~123 million⁷ due to technologies allowing the high-quality, low-cost sequencing of DNA. Machine learning can be used to functionally annotate these sequences, accelerating the discovery of enzymes with useful activities. In the search for a good starting point, ML models can also contribute to filtering the natural diversity of protein catalysts for properties such as stability and solubility. Excitingly, ML can also be applied to generate completely novel protein sequences with the desired function.

Once a protein starting scaffold has been identified, ML models can be applied to help navigate the protein fitness landscape. By training models on experimental data, ML helps prioritize which sets of mutations to test in enzyme engineering campaigns. This approach helps to analyze complex relationships in large datasets, identifying patterns that might be challenging to detect otherwise. This is important, because in most enzyme engineering campaigns only a small fraction of protein sequences can be experimentally sampled. In addition, experimental engineering campaigns tend to focus on single mutational steps, which means ignoring the nonadditive effects of accumulating mutations. ML-assisted directed evolution, however, can be used to predict the fitness of protein variants with several amino acid substitutions. Using this approach, we could for example optimize a halogenase for the late-stage functionalization of the macrolide sorafenin A² and a ketoreductase³, which we engineered for the manufacture of a precursor of the cancer drug ipatasertib.

Stanislav Mazurenko: One immediate application is in addressing the vast number of unannotated sequences in biological databases. With the tremendous variety of biocatalysts already present in nature, we are only scratching the surface with the existing, well-characterized enzymes. The recent breakthroughs in protein structure prediction, such as AlphaFold, have unlocked access to the expansive “structural universe.”⁸ The next major step will be accumulating enough annotated enzyme data to unlock the “functional universe.” ML should be able to give us tools that can predict enzyme activity, substrate scope, co-factors, optimal environments, etc. with high accuracy. We are already increasingly using ML-based annotations in our tools, for example, in EnzymeMiner for automated mining of soluble enzymes⁹. Another area where ML can make an impact is in guiding protein engineering. The complexity of this task is immense, as even a single mutation can completely compromise a protein. We see this very often in our experiments as well as in the data we collected in our databases of mutational effects FireProt^{DB10} and SoluProtMut^{DB11}. Often mutations affecting activity appear far from the active site and act through allostery. The search space for protein engineering, therefore, should span the entire protein, but its combinatorial complexity usually limits rational engineering to just a handful of hotspots. We now already have a range of ML-based tools to navigate the entire search space, from zero-shot predictions generated by foundation protein language models to task-specific predictors or small models purposefully fine-tuned to a mutational dataset for a single enzyme.

Next, quite a few exciting studies on de novo enzyme design have appeared, for instance, suggesting high-affinity protein binders using diffusion models¹² or optimized protein sequences using inverse folding methods¹³. Such models can also be conditioned on a particular geometry of the active site, and there have been attempts to use this conditioning to generate new enzymes. The problem is not solved yet as there are many factors apart from this geometry that define enzymatic activity. However, the field is moving ahead very quickly, so I am curious to see whether the future of enzyme engineering will be dominated by discovering and optimizing existing proteins or designing entirely new ones from scratch. On the other hand, ML is blurring the lines between these two approaches, such as in the cases of de novo

design with the help of an ML model trained on natural sequences.

Finally, AI is being increasingly used in the lab on different levels: hardware control, signal acquisition and processing, data analysis, and design–build–test–learn cycles. I am certain we will see even more exciting AI applications in lab automation as they liberate scientists from repetitive manual tasks and help optimize experimental conditions.

Yang Yang: Due to the vast protein sequence and structure databases, and the similarity between natural language processing and protein sequence/structure processing, machine learning tools, particularly large language models (LLMs), hold the potential to speed up protein engineering and biocatalysis. By discerning the hidden rules in protein fitness landscape, protein LLMs can aid in the navigation of enzyme fitness landscape by focusing on high-fitness regions and by escaping traps of local minima. Generative machine learning (ML) models can potentially allow novel enzyme sequences to be created with good success rate. Furthermore, entirely new enzyme functions may arise using generative enzyme design methods

What are the main challenges in application of machine learning methods to the field of biocatalysis, and how can we overcome them?

Rebecca Buller: Data scarcity and quality remain a significant bottleneck for the application of machine learning in biocatalysis. Experimental datasets are typically small and can be inconsistent, hindering ML models from learning meaningful patterns. Achieving the necessary data quality can be challenging because the generation of large datasets often requires robust and high-throughput assays, which can be complex and resource-intensive to implement. Moreover, the necessity to precisely build and sequence numerous enzyme variants limits the time and cost savings which might be provided by ML. In this context, zero-shot predictors could guide the prediction of protein fitness without any (or less) labeled data from experimental screens. Essentially, zero-shot predictors use general knowledge from large datasets to make accurate predictions about novel or unseen protein variants. Going forward, it will be necessary to determine which zero shot predictors are useful in the context of enzyme function.

Finally, another key challenge is model transferability and generalization. ML models are often trained with data stemming from one protein family using specific substrates and reaction conditions which may not generalize well to others. Potentially, this challenge can be addressed through transfer learning, where models trained in one domain are fine-tuned on smaller, relevant datasets for a new application. In analogy to how chatbots improve through iterative refinement of questions and prompts, protein language models like ProtT5¹⁴, Ankh¹⁵, and ESM2⁸ can be fine-tuned on new data.

Stanislav Mazurenko: One of the main challenges in applying machine learning to biocatalysis is data availability. Enzymatic mechanisms are highly diverse, and, ideally, we want to obtain labeled examples sampled uniformly from this diversity. However, existing data are often sparse and biased, as they are typically collected with specific hypotheses in mind. Sometimes there are ways to account for such biases and adjust algorithms accordingly. Also, approaches like transfer learning and multi-task learning can help leverage knowledge from well-characterized systems to improve predictions for less-explored enzymes. However, all of those have limitations, and often we might not even be aware of particular data biases. Therefore, having hard data collected in a more systematic way in controlled environments will boost ML-based applications. Many high-throughput experimental methods are appearing and have already generated many interesting datasets increasingly being used in ML, but we need more.

Another challenge is data complexity. Enzyme function is influenced by many factors beyond the chemical step, such as stability, solubility, etc. Moreover, every assay has limitations, and researchers often must make assumptions about all those effects when labeling data. If a developer of an ML-based tool is not aware of those factors and cannot account for them in the data preparation step, they will see the data as too noisy for ML and will not use them. Also connected to this point, as many assays cannot isolate specific effects or directly measure the exact labels ML models are trained on, experimental validation remains challenging. Hence, reducing this noise, improving the understanding of experimental limitations and variables, and developing assays for more precise labeling would greatly benefit ML applications.

Lastly, computational infrastructure and expertise can be limiting factors. Many biocatalysis researchers may not have direct access to those required to develop and apply sophisticated ML models. Fostering interdisciplinary collaborations between biochemists, computational biologists, and ML experts may solve this challenge. Additionally, making ML tools more accessible through user-friendly software and platforms can empower researchers to integrate ML into their workflows without needing to become machine learning experts themselves.

Yang Yang: Despite the immense potential of ML methods to accelerate biocatalysis research, many challenges still exist. First, the availability of large biocatalysis datasets, particularly those with a synthetic focus, remains limited. This is particularly true, if one compares these with the deep mutational scanning datasets in other fields of protein science and protein engineering. Nevertheless, with new data collection workflows recently developed to curate large enzyme activity/selectivity-enzyme sequence datasets, within the near future, biocatalysis researchers will likely address this bottleneck. Second, in biocatalysis (and perhaps the generally defined field of catalysis), ML approaches do not always work well with first-principle based approaches, such as the parameterization approach many chemists are undertaking. LLM based description and modeling of proteins have been very successful, but these models are not always easily interpretable. Further studies and collaborations between computational biologists, experimental chemists and computational chemists may address this problem.

How can collaboration between computational experts and biocatalysis experts/experimentalists be fostered?

Rebecca Buller: I believe that the most challenging aspect for interdisciplinary collaborations is to speak a common language. It is important to communicate the power but also the limitations inherent to each technology allowing to better align expectations with achievable outcomes. Forming interdisciplinary teams early on can be key to overcoming communications barriers. For example, in the frame of a large research initiative in Switzerland (National Competence Center for Research – Catalysis), we have involved computational scientists and experimentalists from the early stages of a

project. In this way, we could align the chosen ML approaches more closely with the specific needs of the “wet lab” biocatalysis and vice versa. Additionally, organizing training sessions and workshops—focused on ML for biocatalysis experts and on biocatalysis for ML practitioners—can help establish a shared vocabulary and understanding, ultimately fostering smoother and more productive collaborations.

Stanislav Mazurenko: To foster effective collaboration, educational activities are essential. Both computational experts and experimentalists need to be exposed to each other's fields early in their training. While it is hard to introduce computational experts to the lab bench, they should have at least a basic understanding of protein science and experimental data collection methods to communicate effectively with experimentalists. On the computational side, many researchers are unaware of the vast opportunities in modern biology for applying their skills. Raising awareness about these opportunities will spark more cross-disciplinary collaborations. On the experimental side, biocatalysis experts are usually exposed to some level of programming and data analysis, but what is often lacking are strong data management practices and seeing one's data not only as the ultimate result for testing a single hypothesis but as part of a much larger effort to understand the general rules of life, taking place in labs all over the world. This is a valuable expansion of the perspective on data collection that machine learning brings, and introducing the basics of ML in the curriculum of biochemists has huge potential. From my experience teaching AI to students with no formal training in programming, they are very good at grasping the intuition behind machine learning algorithms and major steps in a typical workflow. However, experimentalists need clear motivation to adopt good data handling practices, including an understanding of how these practices will benefit their work and facilitate collaboration with computational teams, and education is key. Many great initiatives in this respect come from global networks, such as ELIXIR, EBI, NCBI, etc., and we already see results: the groundbreaking success of AlphaFold was possible thanks to all those researchers depositing experimentally determined protein structures in the Protein Data Bank. Even small workshops significantly contribute to raising awareness of the tools and protocols. For example, we constantly get positive

feedback from the Hands-on Computational Enzyme Design Course we organize each year to teach researchers from primarily experimental backgrounds about state-of-the-art computational tools available, including those based on ML.

Yang Yang: Due to the exciting potential of ML-guided enzyme engineering, computational experts and biocatalysis experts have already begun to work closely with each other and many exciting results have emerged over just the past several years. Due to the power of ML approaches in enzyme engineering, many now recognize the potential in this area and numerous exciting collaborations are happening naturally. The US funding agencies see the opportunity in this area, and workshops have been organized to bring together researchers with complementary expertise to brainstorm ideas. For example, I recently participated in an NSF workshop and started a fruitful collaboration with Professor William DeGrado's group in AI-guided enzyme design. It is exciting to see how these projects evolve and there is a lot to expect in the next decade in the field of ML-guided enzyme engineering.

Along similar lines, are there any specific data reporting requirements that would accelerate this collaboration?

Rebecca Buller: Further improvements of ML for biocatalysis will depend on our capacity to generate, curate and store high-quality biological data in the same way as, for example, the community of structural biologists achieve for protein structures in the Protein Data Bank. Beyond the structure-function data, a large body of meta-data describing the experimental conditions will have to be reported in a machine-readable format.

Stanislav Mazurenko: Absolutely! One key paradigm is the FAIR (Findable, Accessible, Interoperable, Reusable) data principles, which aim to make research data more reusable. While it may be unrealistic to expect experimentalists to fully immerse themselves in these technical requirements, I highly recommend depositing data in existing repositories that are already designed to be FAIR. Another important step is adopting the EnzymeML data exchange format¹⁶. This initiative seeks to standardize enzyme kinetic data, which is crucial for managing the complexity of enzymatic mechanisms inherent in biocatalysis. For researchers interested in training their machine learning-based models,

I also recommend following the DOME guidelines, which outline best practices for reporting details about ML pipelines¹⁷. For those who are far from ML but want to improve the way they handle their data, I recommend starting with the so-called “tidy data” approach¹⁸. Tidy datasets make one's life much easier when it comes to data handling, modeling, or visualization, and one does not need any specific background to learn and follow them.

Yang Yang: In the field of directed evolution, historically, researchers have been focusing on beneficial mutations and only report results with improved enzyme variants. To carry out ML-guided enzyme engineering, researchers are now switching to new workflows, where all the variants are sequenced and characterized to learn from effects of deleterious mutations. At this moment, despite the new workflows recently developed, many datasets remain not easily available to the broader biocatalysis community. The community would greatly benefit from the publicly available datasets where attention was paid to uncertainty characterization and statistical analysis.

How do you predict the field will develop in the future?

Rebecca Buller: Catalysis is a complex process often requiring several substrates and cofactors which react in multiple elementary steps. To enable these complex chemical transformations, enzymes may undergo conformational changes. Today, the intricate dynamics of proteins, which often require precise folding and interaction patterns, are difficult to model fully. However, understanding and creating dynamic proteins will be key for more efficient enzyme optimization. Toward this goal, data-driven techniques to analyze molecular dynamic simulations may be applied. Encouragingly, well-curated databases of trajectories are appearing and will play an essential role in training and validating models that consider enhanced representations of protein sequences and structures, incorporating information about 3D conformations, dynamics, and interactions with substrates. Finally, moving toward fully automated closed-loop systems for enzyme engineering, active learning techniques may guide experiments by identifying the most informative enzyme variants to test. The ML models will continuously learn from

experimental feedback, thereby minimizing experimental costs.

Stanislav Mazurenko: In my opinion, we still lack an understanding of the limits of improvement: whether a catalytic activity of a given enzyme, for example, can be pushed to the diffusion limit or if there are intrinsic constraints. Recently, my colleagues have succeeded in improving the stability of an enzyme that had already been stabilized by 23°C through rational protein engineering¹⁹. This raises a question: can such hyperstable enzymes be improved even further? As more data is annotated and integrated into ML models, I am confident ML will help reveal more general patterns and principles, which might elucidate the answer to such questions. In addition to this, ML offers a unique opportunity to test hypotheses on a much larger scale thanks to the increasing availability of annotated data. With the emergence of large datasets like the ESM Metagenomic Atlas²⁰ and ProteinGym²¹, along with automated ML-driven annotations, we can now test our hypothesis using much larger sample sizes. This shift towards data-driven insights will likely lead to more robust conclusions, broader discoveries, and perhaps even the identification of new “rules of biology”, with ML playing a pivotal role in these advancements.

Another promising area is solving enzyme mechanisms. Despite the remarkable progress in technologies for experimental data collection, conducting experiments and analyzing activity data to unravel individual catalytic steps is still a long and challenging process. If ML can help accelerate this process – by guiding the optimal design of experiments, data fitting, and interpretation, for example – it could significantly speed up enzyme characterization and make engineering more focused. More annotated data will help us better predict enzyme activity and its rate-limiting steps. Similarly, modifying substrate specificity or enantioselectivity remains a highly desirable yet elusive goal. Several promising studies have explored the fusion of machine learning and directed evolution to address those challenges. I am particularly curious to discover if ML can further reduce the need for extensive experimental data collection, for instance, thanks to the further development of zero-shot and low-N prediction methods.

Lastly, recent papers in protein science have started using quantum machine learning to make predictions²². Although we have yet to

witness quantum advantage over classical ML, quantum computing is rapidly becoming a reality. As this technology is entirely new to Life Sciences, we should begin considering how to prepare future generations to harness the potential of quantum computing.

Yang Yang: We are at a very exciting time for ML-guided enzyme research. ML tools have been tremendously successful in protein structure prediction and more recently, in protein design. Translation of these tools to enzyme research is likely transformative in many aspects. In the traditional workflow of directed enzyme evolution, only the top variants are sequenced and carried forward for further engineering. In ML-guided directed evolution (MLDE), with new technologies to significantly bring down the costs of sequencing and higher throughput enzyme activity assays, our community is in a new era to generate large biocatalysis datasets. Through the integration of computational biology and enzyme engineering, we might have a broadly applicable and easily affordable workflow of MLDE for all biocatalysis researchers to use in daily research. Developing and applying enzyme specific LLMs may lead to powerful new methodologies in enzyme engineering. Integrating chemical knowledge with computational methods will significantly push the boundary of the field.

The interview was conducted by Dr. Majda Bratovič.

Published online: 02 October 2025

References

- Patsch, D. et al. Enriching productive mutational paths accelerates enzyme evolution. *Nat. Chem. Biol.* **20**, 1662–1669 (2024).
- Büchler, J. et al. Algorithm-aided engineering of aliphatic halogenase WelO5* for the asymmetric late-stage functionalization of soraphens. *Nat. Commun.* **13**, 371 (2022).
- Honda Malca, S. et al. Effective engineering of a ketoreductase for the biocatalytic synthesis of an ipatasertib precursor. *Commun. Chem.* **7**, 1–11 (2024).
- Protein Engineering Portal | Loschmidt Laboratories. <https://loschmidt.chemi.muni.cz/portal/>.
- Richardson, L. et al. MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2023).
- The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
- Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**, 2542 (2018).
- Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
- Hon, J. et al. EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities. *Nucleic Acids Res.* **48**, W104–W109 (2020).
- Stourac, J. et al. FireProtDB: database of manually curated protein stability data. *Nucleic Acids Res.* **49**, D319–D324 (2021).
- Velecký, J. et al. SoluProtMutDB: a manually curated database of protein solubility changes upon mutations. *Comput. Struct. Biotechnol. J.* **20**, 6339–6347 (2022).
- Watson, J. L. et al. De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
- Sumida, K. H. et al. Improving protein expression, stability, and function with ProteinMPNN. *J. Am. Chem. Soc.* **146**, 2054–2061 (2024).
- Elnaggar, A. et al. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2022).
- Elnaggar, A. et al. Ankh: optimized protein language model unlocks general-purpose modelling. Preprint at <https://doi.org/10.48550/arXiv.2301.06568> (2023).
- Lauterbach, S. et al. EnzymeML: seamless data flow and modeling of enzymatic data. *Nat. Methods* **20**, 400–402 (2023).
- Walsh, I. et al. DOME: recommendations for supervised machine learning validation in biology. *Nat. Methods* **18**, 1122–1127 (2021).
- Wickham, H. Tidy data. *J. Stat. Softw.* **59**, 1–23 (2014).
- Kunka, A. et al. Advancing enzyme's stability and catalytic efficiency through synergy of force-field calculations, evolutionary analysis, and machine learning. *ACS Catal.* **13**, 12506–12518 (2023).
- ESM Metagenomic Atlas | Meta AI. <https://esmatlas.com>.
- Notin, P. et al. ProteinGym: large-scale benchmarks for protein fitness prediction and design. *Adv. Neural Inf. Process. Syst.* **36**, 64331–64379 (2023).
- Natęcz-Charkiewicz, K., Charkiewicz, K. & Nowak, R. M. Quantum computing in bioinformatics: a systematic review mapping. *Brief. Bioinform.* **25**, bbae391 (2024).

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© Springer Nature Limited 2025