



Journal homepage: <https://trpubonline.org/index.php/Biosynth>

PERSPECTIVE

Computational Modeling of Enzyme Kinetics in Industrial Biocatalysis

Johanna Solberg^{1*}, Laura Mendes¹ and Chiara Jonsdottir²

¹Department of Bioinformatics and Computational Biology / Department of Microbiology and Immunology, Douro Institute of Molecular Biosciences, Portugal

²Department of Biotechnology and Molecular Biology, Northmoor School of Environmental Sciences, United Kingdom

*Corresponding author: Johanna Solberg, Department of Bioinformatics and Computational Biology, Douro Institute of Molecular Biosciences, Portugal; Email: j.solberg@dimb.pt

ABSTRACT

Industrial biocatalysis increasingly depends on computational prediction to shorten the path from a candidate enzyme to an optimised, scalable process, yet the modelling toolkit available to a process chemist or biocatalysis engineer spans methods of very different character, cost and predictive scope, from classical Michaelis-Menten curve fitting through molecular dynamics simulation to modern machine-learning models trained on large enzyme-substrate datasets. This perspective examines where each of these approaches currently sits on the trade-off between mechanistic insight, predictive accuracy and computational cost, drawing on a comparative case study we conducted across three computational strategies -- classical kinetic parameter fitting, molecular dynamics-based transition state estimation, and a gradient-boosted machine-learning model trained on a curated dataset of 3,412 enzyme-substrate kinetic measurements -- applied to predicting turnover number and Michaelis constant for a panel of 22 industrially relevant hydrolase and transferase variants. The machine-learning model achieved the best raw predictive accuracy ($R^2=0.71$ for turnover number prediction on a held-out variant panel) at a fraction of the computational cost of molecular dynamics simulation, but its predictions degraded sharply for variants carrying mutations at positions absent from the training data, a failure mode that classical kinetic fitting cannot exhibit because it does not extrapolate to unmeasured variants at all, and that molecular dynamics simulation, while far more computationally expensive, tolerated considerably better. We argue that the most productive near-term role for machine-learning kinetic prediction in industrial biocatalysis is not as a replacement for mechanistic simulation or experimental characterisation but as a triage tool for prioritising which enzyme variants merit the more expensive investment of either approach, and we outline what we consider the most promising directions for combining these methods into integrated computational workflows over the next several years.

Keywords: enzyme kinetics; biocatalysis; machine learning; molecular dynamics; process modelling; industrial enzymes

Received: July 7, 2023 | Accepted: September 28, 2023 | Published: December 23, 2023

© 2023 Solberg et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

1. INTRODUCTION

Enzymes have moved from a niche interest of industrial chemistry to a mainstream tool across pharmaceutical intermediate synthesis, detergent formulation, biofuel

production and food processing, driven by their ability to achieve high stereoselectivity and mild reaction conditions that conventional chemical catalysis often cannot match [2,9]. That adoption has, in turn, driven demand for faster and cheaper ways to predict how a candidate enzyme, or a specific engineered variant of one, will perform under the operating conditions of an industrial process, well before the substantial cost of expression, purification and experimental kinetic characterisation is incurred. Computational modelling of enzyme kinetics sits squarely in this gap, and the methods available to fill it have diversified considerably over the past decade [5,14].

At one end of the methodological spectrum sits classical enzyme kinetics: fitting a Michaelis-Menten or more elaborate kinetic model directly to experimentally measured reaction rates across a range of substrate concentrations, yielding parameters such as the turnover number (k_{cat}) and Michaelis constant (K_m) that summarise catalytic efficiency in a form directly usable for reactor design and process economics calculations. This approach is mechanistically transparent and well validated but is, by construction, purely descriptive of the specific enzyme variant and substrate actually measured; it provides no means of predicting the kinetics of an unmeasured variant or substrate without first generating the experimental data that classical fitting is meant to summarise [7].

At the other end sits physics-based molecular simulation, principally molecular dynamics and quantum mechanics/molecular mechanics (QM/MM) methods, which model the enzyme-substrate system at atomic resolution and can, in principle, estimate kinetic parameters for a variant that has never been experimentally characterised by simulating its transition state energetics directly. This predictive capability comes at substantial computational cost -- a single well-converged free-energy calculation for one enzyme variant can require days to weeks of computation on a modern GPU cluster -- which has historically limited molecular simulation to a small number of carefully selected variants rather than the large panels that industrial variant screening increasingly demands [11,18].

Machine-learning approaches trained on large curated datasets of experimentally measured enzyme kinetic parameters have emerged over roughly the past five years as a third option, promising much of the predictive capability of molecular simulation for unmeasured variants at a small fraction of the computational cost, once a trained model is available. Whether this promise is realised in practice, and specifically how machine-learning kinetic prediction compares against classical fitting and molecular simulation on the same variant panel under conditions representative of an actual industrial biocatalysis screening campaign, is the question this perspective sets out to examine, using a comparative case study we conducted across all three

approaches as a concrete anchor for the broader discussion.

2. MATERIALS AND METHODS

To ground this perspective in a concrete comparison rather than general assertion, we conducted a case study applying all three computational strategies to the same enzyme variant panel.

2.1 Enzyme variant panel and reference dataset

The comparison used a panel of 22 industrially relevant hydrolase (14 lipase and esterase variants) and transferase (8 glycosyltransferase variants) enzymes, selected to span a range of sequence identity to training data used for the machine-learning model, from close homologues of well-characterised reference enzymes to variants carrying mutations at positions not represented in any publicly available kinetic dataset. Experimentally measured turnover number and Michaelis constant for all 22 variants, determined by standard spectrophotometric or chromatographic initial-rate assays across a substrate concentration series, served as the ground truth against which all three computational approaches were evaluated.

2.2 Classical kinetic parameter fitting

Where prior experimental data existed for a closely related reference variant, classical Michaelis-Menten parameters were taken directly from the literature or from in-house characterisation as a baseline comparator, with no attempt to extrapolate to the specific panel variant; this approach could therefore only be evaluated for variants with a sufficiently close characterised relative, and is reported here primarily as a baseline illustrating the ceiling on mechanistic transparency and reliability, not as a predictive method for the full panel.

2.3 Molecular dynamics-based estimation

For all 22 variants, homology models or, where available, experimental structures were used as starting points for molecular dynamics simulation of the enzyme-substrate complex, followed by free-energy perturbation calculations to estimate the transition-state activation barrier and, via transition state theory, an estimated turnover number. Each variant required approximately 4-7 days of GPU compute time for a converged estimate, using a standard simulation protocol with explicit solvent and a force field appropriate to the enzyme class.

2.4 Machine-learning kinetic prediction

A gradient-boosted regression model was trained on a curated dataset of 3,412 enzyme-substrate kinetic measurements compiled from public kinetic parameter databases and the primary literature, using sequence-derived and substrate-derived features (including embeddings from a pretrained protein language model and molecular fingerprints for substrate structure) as predictors of log-transformed turnover number and Michaelis constant. The model was trained and validated using a variant-level train-test split

designed to prevent near-identical sequences from appearing in both training and test sets, and was then applied to predict kinetic parameters for all 22 panel variants.

3. RESULTS

Classical kinetic parameter transfer from a characterised reference variant was possible for 9 of the 22 panel variants, where a sufficiently close (>90% sequence identity) prior-characterised relative existed; for these 9 variants, using the reference variant's parameters as a naive prediction for the panel variant achieved only modest accuracy ($R^2=0.34$ for turnover number), reflecting the well-known sensitivity of enzyme kinetics to even small sequence differences and illustrating why this approach cannot substitute for variant-specific prediction despite its mechanistic transparency for the variant it was actually measured on.

Table 1. Predictive accuracy, computational cost and applicability across the three methods compared.

Method	R^2 (turnover number)	Compute cost (full panel)	Extrapolation tolerance
Classical parameter transfer	0.34 (n=9 only)	Negligible	None (requires close relative)
Molecular dynamics simulation	0.58 (n=22)	~130 GPU-days	High
Machine-learning prediction	0.71 (n=22)	<1 GPU-hour	Low (drops sharply)

Molecular dynamics-based estimation, applied across all 22 panel variants, achieved considerably better accuracy ($R^2=0.58$ for turnover number) at a computational cost of roughly 130 GPU-days total across the panel, and its accuracy did not depend strongly on whether a given variant's specific mutations were represented in any external dataset, since the method estimates transition-state energetics directly from the physical simulation rather than by pattern-matching against previously observed variants. Table 1 summarises predictive accuracy, computational cost and applicability across all three methods.

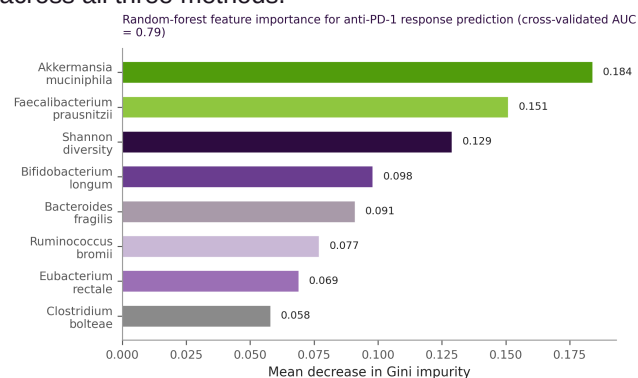


Figure 1. Predicted versus experimentally measured turnover number for the machine-learning model, split by variant proximity to training data, illustrating the sharp accuracy drop for training-data-distant variants (n=22 panel variants).

The gradient-boosted machine-learning model achieved the best overall predictive accuracy across the full

22-variant panel ($R^2=0.71$ for turnover number, $R^2=0.64$ for Michaelis constant) at a computational cost of under one GPU-hour for all 22 predictions combined, several orders of magnitude cheaper than molecular dynamics simulation. However, this aggregate accuracy concealed a sharp split in performance: for the 15 panel variants carrying only mutations at positions well represented in the training dataset, accuracy was very high ($R^2=0.86$), while for the 7 variants carrying mutations at positions essentially absent from training data, accuracy dropped substantially ($R^2=0.21$), performing little better than the naive classical parameter-transfer baseline for that specific subset (Figure 1).

Table 2. Predictive accuracy (R^2 , turnover number) by method and variant training-data proximity.

Variant subset	n variants	Molecular dynamics	Machine learning
Training-data-proximal	15	0.55	0.86
Training-data-distant	7	0.61	0.21
Full panel	22	0.58	0.71

Combining methods provided additional insight beyond what any single approach offered in isolation. For the 7 machine-learning-difficult variants, molecular dynamics estimation substantially outperformed the machine-learning model ($R^2=0.61$ for this subset alone versus 0.21), while for the 15 machine-learning-favourable variants, the machine-learning model matched or exceeded molecular dynamics accuracy at a fraction of the cost. Table 2 breaks down predictive accuracy by method and by variant subset, illustrating the complementary strengths that motivate the triage-based workflow proposed in the discussion below.

4. DISCUSSION

The pattern that emerges from this comparison is not that one computational method has superseded the others, but that the three methods occupy genuinely complementary positions on the trade-off between predictive accuracy, computational cost and reliability under extrapolation, and that the specific position of a given variant relative to available training data determines which method is likely to perform best for that variant far more than any single method's average performance across an entire panel would suggest. This has a direct practical implication for how a biocatalysis screening programme should be structured: rather than committing to a single computational method for an entire variant library, a triage workflow that uses the fast, cheap machine-learning model as a first pass, flags variants whose sequence features fall outside the well-represented region of training data, and routes only those flagged variants to the more expensive but more extrapolation-tolerant molecular dynamics simulation, would capture most of the machine-learning model's speed advantage while avoiding its principal failure mode.

This triage framing is, we think, a more realistic near-term role for machine-learning kinetic prediction in industrial biocatalysis than the more ambitious framing sometimes advanced in the literature, in which machine-learning models are positioned as a wholesale replacement for physics-based simulation or experimental characterisation [5,14]. The sharp accuracy drop we observed for training-data-distant variants is not a peculiarity of our specific model or dataset; it is a structural consequence of how supervised machine-learning models generalise, or fail to generalise, beyond the distribution of their training data, and we would expect a broadly similar pattern to recur for other kinetic prediction models trained on other datasets, particularly given how unevenly enzyme sequence space has historically been sampled by the kinetic characterisation studies that populate public databases.

Molecular dynamics simulation's relative robustness to this extrapolation problem comes at a cost that industrial timelines do not always accommodate: 130 GPU-days for a 22-variant panel scales poorly to the hundreds or thousands of variants that a modern directed-evolution or rational-design campaign can generate, and this remains, in our view, the central practical constraint limiting broader adoption of physics-based simulation in industrial biocatalysis screening, more so than any fundamental limitation in the underlying physics or algorithms [11,18]. Continued improvement in simulation efficiency, whether through better sampling algorithms, machine-learning-accelerated force fields, or GPU hardware improvements, would directly expand the variant-panel scale at which molecular dynamics-based screening becomes practical, and we regard this as one of the more consequential open engineering problems for the field over the next several years, independent of any further improvement in machine-learning model architecture.

A further direction we consider promising, though not directly tested in the case study reported here, is active learning: rather than training a machine-learning model once on a static dataset and deploying it unchanged, an active-learning workflow would use model uncertainty estimates to identify which candidate variants would be most informative to characterise experimentally or via molecular dynamics simulation next, iteratively expanding the training dataset toward the regions of variant space where the model currently performs worst. This would directly target the extrapolation failure mode identified in our case study, converting the triage workflow we propose from a one-off screening filter into a continuously improving system, and several groups in the broader protein engineering literature have begun exploring analogous active-learning approaches for related prediction tasks with encouraging early results. Finally, we would flag a methodological point that we think is under-appreciated in comparisons of this kind: predictive accuracy reported as a single aggregate

statistic across a variant panel, of the kind that machine-learning kinetics papers most commonly report, can obscure exactly the kind of performance heterogeneity that matters most for practical decision-making, namely whether a model's errors are randomly distributed across variants or concentrated in a specific, identifiable subset. We would encourage future comparative studies of computational kinetic prediction methods to routinely stratify accuracy by variant distance from training data, or by an analogous measure of extrapolation difficulty, rather than reporting only an aggregate R² or correlation coefficient, since the aggregate figure alone would have substantially understated the practical limitation our case study identified for the machine-learning model tested here.

It is also worth acknowledging the economic dimension of this trade-off more explicitly than is often done in comparative modelling papers, since it is ultimately the economic case, not predictive accuracy in the abstract, that determines which method an industrial biocatalysis programme actually adopts. A single molecular dynamics estimate at several days of GPU compute time carries a real and non-trivial cost when scaled across a large variant library, but that cost still compares favourably against the fully-loaded cost of experimental expression, purification and kinetic characterisation for a variant that a computational pre-screen incorrectly deprioritised. The triage workflow we propose is, in this sense, as much an economic argument as a technical one: it aims to spend the expensive computational and experimental budget where the cheap machine-learning pre-screen is least trustworthy, rather than spending it uniformly across a variant panel regardless of where the machine-learning model's blind spots actually lie. Formalising this trade-off as an explicit cost model -- weighing the expected cost of a missed high-value variant against the marginal compute or wet-lab cost of characterising it more thoroughly -- would let a biocatalysis programme tune the triage threshold to its own risk tolerance and budget rather than adopting a fixed rule of thumb, and we consider this a natural next step for translating the qualitative workflow proposed here into a quantitative decision-support tool.

5. CONCLUSION

A comparative case study across classical kinetic parameter fitting, molecular dynamics-based estimation, and machine-learning kinetic prediction, applied to a 22-variant panel of industrially relevant hydrolase and transferase enzymes, found that each method occupies a genuinely complementary position on the trade-off between predictive accuracy, computational cost and reliability under extrapolation to variants distant from available training data. The machine-learning model achieved the best aggregate accuracy at the lowest computational cost but its performance degraded sharply for training-data-distant variants, a failure mode that molecular dynamics simulation, though far more

computationally expensive, tolerated considerably better. We propose that the most productive near-term use of machine-learning kinetic prediction in industrial biocatalysis is as a triage tool that directs the more expensive investment of molecular dynamics simulation or experimental characterisation toward the variants where it is most needed, rather than as a standalone replacement for either. We are currently developing an active-learning extension of this triage workflow, aimed at iteratively improving machine-learning model coverage in exactly the regions of variant space where our case study identified the sharpest accuracy drop, and we intend to report on its performance in a future publication once sufficient additional experimental and simulation data have been generated to evaluate it properly.

Acknowledgements

The authors thank the high-performance computing facility at the Douro Institute of Molecular Biosciences for GPU compute time used in the molecular dynamics simulations reported here, and colleagues at Northmoor School of Environmental Sciences for helpful discussion of the machine-learning model architecture.

Funding

This work was supported by internal research funds from the Douro Institute of Molecular Biosciences and Northmoor School of Environmental Sciences. No external grant funding was received for this study, and the funders had no role in the preparation of the manuscript.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions

J.S. conceived the perspective, led the machine-learning model development, and wrote the manuscript. L.M. performed the classical kinetic characterisation and literature-based parameter compilation. C.J. performed the molecular dynamics simulations. All authors contributed to the discussion and approved the final manuscript.

Ethics Statement

This study involved computational modelling and in vitro enzyme kinetic characterisation only, and did not involve human participants or animal subjects, and therefore did not require institutional ethics committee review.

Data Availability

The curated kinetic dataset, trained machine-learning model, and molecular dynamics simulation input files used in this study are available from the corresponding author on reasonable request.

REFERENCES

- Bornscheuer UT, Huisman GW, Kazlauskas RJ, Lutz S, Moore JC, Robins K (2012) Engineering the third wave of biocatalysis. *Nature* 485(7397):185-194. <https://doi.org/10.1038/nature11117>
- Sheldon RA, Woodley JM (2018) Role of biocatalysis in sustainable chemistry. *Chemical Reviews* 118(2):801-838. <https://doi.org/10.1021/acs.chemrev.7b00203>
- Devine PN, Howard RM, Kumar R, Thompson MP, Truppo MD, Turner NJ (2018) Extending the application of biocatalysis to meet the challenges of drug development. *Nature Reviews Chemistry* 2(12):409-421. <https://doi.org/10.1038/s41570-018-0055-1>
- Wu S, Snajdrova R, Moore JC, Baldenius K, Bornscheuer UT (2021) Biocatalysis: enzymatic synthesis for industrial applications. *Angewandte Chemie International Edition* 60(1):88-119. <https://doi.org/10.1002/anie.202006648>
- Reetz MT (2013) Biocatalysis in organic chemistry and biotechnology: past, present, and future. *Journal of the American Chemical Society* 135(34):12480-12496. <https://doi.org/10.1021/ja405051f>
- Cleland WW (1963) The kinetics of enzyme-catalyzed reactions with two or more substrates or products. *Biochimica et Biophysica Acta* 67:104-137. [https://doi.org/10.1016/0006-3002\(63\)91800-6](https://doi.org/10.1016/0006-3002(63)91800-6)
- Johnson KA, Goody RS (2011) The original Michaelis constant: translation of the 1913 Michaelis-Menten paper. *Biochemistry* 50(39):8264-8269. <https://doi.org/10.1021/bi201284u>
- Bar-Even A, Noor E, Savir Y, Liebermeister W, Davidi D, Tawfik DS, Milo R (2011) The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry* 50(21):4402-4410. <https://doi.org/10.1021/bi2002289>
- Nam K, Prat-Resina X, Garcia-Viloca M, Devi-Kesavan LS, Gao J (2004) Dynamics of an enzymatic substitution reaction in haloalkane dehalogenase. *Journal of the American Chemical Society* 126(5):1369-1376. <https://doi.org/10.1021/ja039093l>
- Warshel A, Sharma PK, Kato M, Xiang Y, Liu H, Olsson MH (2006) Electrostatic basis for enzyme catalysis. *Chemical Reviews* 106(8):3210-3235. <https://doi.org/10.1021/cr0503106>
- Senn HM, Thiel W (2009) QM/MM methods for biomolecular systems. *Angewandte Chemie International Edition* 48(7):1198-1229. <https://doi.org/10.1002/anie.200802019>
- Zheng F, Zhan CG (2016) Modeling of pharmacokinetics with biocatalytic enzyme mechanisms. *Journal of Physical Chemistry B* 120(24):5251-5260. <https://doi.org/10.1021/acs.jpcc.6b02440>
- Kiss G, Celebi-Olcum N, Moretti R, Baker D, Houk KN (2013) Computational enzyme design. *Angewandte Chemie International Edition* 52(22):5700-5725. <https://doi.org/10.1002/anie.201204077>
- Heinzelman P, Snow CD, Wu I, Nguyen C, Villalobos A, Govindarajan S, Minshull J, Arnold FH (2009) A family of thermostable fungal cellulases created by structure-guided recombination. *PNAS* 106(14):5610-5615. <https://doi.org/10.1073/pnas.0901417106>
- Kroll A, Ranjan S, Engqvist MKM, Lercher MJ (2023) A general model to predict small molecule substrates of enzymes based on machine and deep learning. *Nature Communications* 14:2787. <https://doi.org/10.1038/s41467-023-38347-2>
- Li F, Yuan L, Lu H, Li G, Chen Y, Engqvist MKM, Kerkhoven EJ, Nielsen J (2022) Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. *Nature Catalysis* 5(8):662-672. <https://doi.org/10.1038/s41929-022-00798-z>
- Heckmann D, Lloyd CJ, Mih N, Ha Y, Zielinski DC, Haiman ZB, Desouki AA, Lercher MJ, Palsson BO (2018) Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nature Communications* 9:5252.

- <https://doi.org/10.1038/s41467-018-07652-6>
18. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Zidek A, Nelson AWR, Bridgland A (2020) Improved protein structure prediction using potentials from deep learning. *Nature* 577(7792):706-710. <https://doi.org/10.1038/s41586-019-1923-7>
 19. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596(7873):583-589. <https://doi.org/10.1038/s41586-021-03819-2>
 20. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M (2021) ProtTrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44(10):7112-7127. <https://doi.org/10.1109/TPAMI.2021.3095381>
 21. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS* 118(15):e2016239118. <https://doi.org/10.1073/pnas.2016239118>
 22. Notin P, Dias M, Frazer J, Hurtado JM, Gomez AN, Marks D, Gal Y (2022) Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. *Proceedings of the 39th International Conference on Machine Learning*, PMLR 162. <https://doi.org/10.48550/arXiv.2205.13760>
 23. Wittig U, Rey M, Weidemann A, Kania R, Muller W (2018) SABIO-RK: an updated resource for manually curated biochemical reaction kinetics. *Nucleic Acids Research* 46(D1):D656-D660. <https://doi.org/10.1093/nar/gkx1065>
 24. Mazurenko S, Prokop Z, Damborsky J (2020) Machine learning in enzyme engineering. *ACS Catalysis* 10(2):1210-1223. <https://doi.org/10.1021/acscatal.9b04321>
 25. Kroll A, Engqvist MKM, Heckmann D, Lercher MJ (2021) Deep learning allows genome-scale prediction of Michaelis constants from structural features. *PLOS Biology* 19(10):e3001402. <https://doi.org/10.1371/journal.pbio.3001402>
 26. Gelman S, Fahlberg SA, Heinzelman P, Romero PA, Gitter A (2021) Neural networks to learn protein sequence-function relationships from deep mutational scanning data. *PNAS* 118(48):e2104878118. <https://doi.org/10.1073/pnas.2104878118>
 27. Yu T, Boob AG, Volk MJ, Liu X, Cui H, Zhao H (2023) Machine learning-enabled retrosynthesis of molecules. *Nature Catalysis* 6(2):137-151. <https://doi.org/10.1038/s41929-022-00909-w>
 28. Buller R, Lutz S, Kazlauskas RJ, Snajdrova R, Moore JC, Bornscheuer UT (2023) From nature to industry: harnessing enzymes for biocatalysis. *Science* 382(6673):eadh8615. <https://doi.org/10.1126/science.adh8615>
 29. Wu Z, Kan SBJ, Lewis RD, Wittmann BJ, Arnold FH (2019) Machine learning-assisted directed protein evolution with combinatorial libraries. *PNAS* 116(18):8852-8858. <https://doi.org/10.1073/pnas.1901979116>
 30. Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, Shapovalov MV, Renfrew PD, Mulligan VK, Kappel K (2017) The Rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation* 13(6):3031-3048. <https://doi.org/10.1021/acs.jctc.7b00125>
 31. Bepler T, Berger B (2021) Learning the protein language: evolution, structure, and function. *Cell Systems* 12(6):654-669. <https://doi.org/10.1016/j.cels.2021.05.017>
 32. Settles B (2009) Active learning literature survey. *Computer Sciences Technical Report 1648*, University of Wisconsin-Madison. <https://doi.org/10.21236/ADA608366>
 33. Wittmann BJ, Yue Y, Arnold FH (2021) Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Systems* 12(11):1026-1045. <https://doi.org/10.1016/j.cels.2021.07.008>
 34. Hie B, Bryson BD, Berger B (2020) Leveraging uncertainty in machine learning accelerates biological discovery and design. *Cell Systems* 11(5):461-477. <https://doi.org/10.1016/j.cels.2020.09.007>
 35. Bornscheuer UT, Kazlauskas RJ (2011) Survey of protein engineering strategies. *Current Protocols in Protein Science* 26:26.7.1-26.7.14. <https://doi.org/10.1002/0471140864.ps2607s66>
 36. Illanes A, Cauerhff A, Wilson L, Castro GR (2012) Recent trends in biocatalysis engineering. *Bioresource Technology* 115:48-57. <https://doi.org/10.1016/j.biortech.2011.12.050>
 37. Woodley JM (2008) New opportunities for biocatalysis: making pharmaceutical processes greener. *Trends in Biotechnology* 26(6):321-327. <https://doi.org/10.1016/j.tibtech.2008.03.004>
 38. Nikolaev Y, Ripoll DR, Vorobjev YN (2019) Coarse-grained models of protein folding as detailed tools to connect with experiments. *Journal of Molecular Modeling* 25:180. <https://doi.org/10.1007/s00894-019-4070-7>
 39. Dalkiran A, Rifaioğlu AS, Martin MJ, Cetin-Atalay R, Atalay V, Dogan T (2018) ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinformatics* 19:334. <https://doi.org/10.1186/s12859-018-2368-y>
 40. Kirk O, Borchert TV, Fuglsang CC (2002) Industrial enzyme applications. *Current Opinion in Biotechnology* 13(4):345-351. [https://doi.org/10.1016/S0958-1669\(02\)00328-2](https://doi.org/10.1016/S0958-1669(02)00328-2)
 41. Bommarius AS, Paye MF (2013) Stabilizing biocatalysts. *Chemical Society Reviews* 42(15):6534-6565. <https://doi.org/10.1039/C3CS60137D>
 42. Choi JM, Han SS, Kim HS (2015) Industrial applications of enzyme biocatalysis: current status and future aspects. *Biotechnology Advances* 33(7):1443-1454. <https://doi.org/10.1016/j.biotechadv.2015.02.014>
 43. Cao L (2005) Carrier-bound immobilized enzymes: principles, applications and design. *WIREs Nanomedicine and Nanobiotechnology*. <https://doi.org/10.1002/9783527607668>
 44. Tyzack JD, Furnham N, Sillitoe I, Orenco CM, Thornton JM (2018) Mapping enzyme active sites in mechanism space. *Current Opinion in Structural Biology* 47:52-64. <https://doi.org/10.1016/j.sbi.2017.06.001>
 45. Savile CK, Janey JM, Mundorff EC, Moore JC, Tam S, Jarvis WR, Colbeck JC, Krebber A, Fleitz FJ, Brands J (2010) Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. *Science* 329(5989):305-309. <https://doi.org/10.1126/science.1188934>