ACS Catalysis

pubs.acs.org/acscatalysis
Viewpoint

Mobilizing the Biocatalysis Community for Reproducible and Reusable Data Collection

Sérgio M. Marques, Joan Planas-Iglesias, Jan Velecký, Milos Musil, Yasuhisa Asano, Tomasz Borowski, Vânia Brissos, Marco Cespugli, Koar Chorozian, Mohammad Dadashipour, Elif Erdem, Erica Elisa Ferrandi, Konstantinos Grigorakis, Anna Kluza, Janina Lawniczek, Konstantinos Makryniotis, Daniela Monti, Bettina Nestl, Anna C. Ngo, Efstratios Nikolaivits, Stefania Patti, Christina Pentari, Carolina F. Rodrigues, Tobias Schopper, Karolina Seweryn-Ożóg, Maciej Szaleniec, André Taborda, Mateusz Tataruch, Dirk Tischler, Evangelos Topakas, Jingyu Wang, Patrycja Wójcik, Agnieszka M. Wojtkiewicz, John M. Woodley, Olga Zastawny, Lígia O. Martins, Marco Fraaije, Jürgen Pleiss, Santiago Schnell, Jiri Damborsky,* Stanislav Mazurenko,* and David Bednar*

Cite This: *ACS Catal.* 2026, 16, 8858−8868

INTRODUCTION

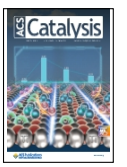
Importance of Sharing Experimental Data

Science is an ever-evolving endeavor, with all new research grounded in knowledge gained in previous studies and publications. This applies not only at the level of theory and fundamental knowledge, but also at the level of specific data. In the context of enzyme research, that includes information on properties such as protein production and folding, protein solubility, stability, catalytic activity, together with specificity and stereoselectivity, as well as regulatory effects as activation and inhibition, and kinetics, which are crucial for multiple practical reasons. In the fields of biology and biochemistry, the availability of high-quality experimental data has already contributed to several breakthroughs over time. One example is AlphaFold 2,[1] released in 2021, a machine learning-based tool that predicts the 3D structures of proteins with unprecedented accuracy. Its release represented a major breakthrough in structural biology, addressing a long-standing challenge that had persisted for decades. A key element in the success of AlphaFold was the large number of experimental protein structures available in the Protein Data Bank (ca. 159,000 in 2019).[2] This was made possible because the deposition of crystallographic, nuclear magnetic resonance (NMR), and electron microscopy (cryo-EM) structures in a uniform format into databases became the gold standard and a strict requirement for their publication three decades before the AlphaFold release.[3,4] Thanks to the high quality and the large volume of its data, the Protein Data Bank also enabled the development of molecular docking and other tools. Other examples are UniProt[5] and BRENDA,[6] databases that contributed to functional prediction tools,[7−9] metabolic modeling,[10−12] and large-scale enzyme design efforts.[13−16] Their success relies heavily on community contributions, data quality checks, and manual curation.

Bottleneck in Enzyme Engineering

Enzyme engineering and predictive biotechnology still face numerous challenges.[17] Predicting enzyme activity, selectivity, stability, and solubility remains difficult, not only because of the complexity of the underlying physical processes, but also due to the limited availability and heterogeneity of high-quality experimental data. The fast-evolving machine-learning techniques require training on reliable data to generate accurate predictions. Traditional low-throughput methods often offer greater accuracy, reproducibility, and interpretability compared to current high-throughput techniques, and their contribution remains highly valuable. Most research publications report experimental results in the form of tables and figures. Naturally, these results need to be presented in a form understandable to humans who will read them. However, search algorithms often miss such results, because they are either not available in a machine-readable format or they are hidden in the Supporting materials, which can be even harder to trace. Although several high-quality repositories exist, namely STRENDA DB[18] and SABIO-RK[19] for kinetic measurements, BRENDA[6] for enzyme functional annotations, and domain-specific resources such as FireProt^{DB} and SoluProtMut^{DB} for stability and solubility, respectively, there is currently no universally mandated deposition venue across journals and no uniform reporting practice for stability and solubility data.[20] This fragmentation leads to heterogeneous metadata and inconsistency in unit conventions and uncertainty reporting. Our recommendations, therefore, aim

Received: November 10, 2025
Revised: March 24, 2026
Accepted: March 25, 2026
Published: May 5, 2026



https://doi.org/10.1021/acscatal.5c07904
ACS Catal. 2026, 16, 8858−8868

to articulate a coordinated, interoperable pathway that bridges kinetic, stability, and solubility measurements under shared formats and vocabularies. This clearly represents a bottleneck in the development of future tools devoted to engineering better biocatalysts. The situation is gradually improving due to increased community awareness of the need for open science and the publication of data that adheres to the FAIR Data Principles (see below). High-quality data repositories can only emerge from community-wide agreement on how enzyme data are reported. This entails: (i) consensus on standardized reporting across disciplines and journals, and (ii) widespread adoption of established author guidelines for enzymology and biocatalysis, such as the STRENDA guidelines, now embedded in the author instructions of 55 peer-reviewed biochemistry journals.[18]

Negative Data Matters

Apart from general data availability, to generate accurate predictive models, it is necessary to have balanced datasets containing both positive and negative results. Otherwise, the models will be biased and inaccurate. It is important to note that in machine learning, data labelling as positive or negative is subjective, typically based on the problem context: the positive class usually represents the condition of interest, while the negative class represents its absence. In what follows, we will refer to "negative results" in the context of biocatalysis: the subset of experimental conditions or protein variants tested that did not lead to the desired outcome. Some examples include protein variants that show a deleterious effect on the desired objective, such as catalytic activity, those that have no detectable effect within statistical limits relative to the reference (typically wild type), or those that could not be characterized due to failed expression or purification steps. This notion is often subjective and experiment-specific: what is an undesirable outcome in one experiment (e.g., compromised solubility and yield) can be considered a success in another (e.g., formation of catalytically active inclusion bodies to facilitate enzyme purification)[21] or change significantly in different experimental conditions (e.g., optimized expression). While positive data are available across the literature, the negative or unsuccessful results are often left unpublished.[22] This type of data is highly valuable for training robust predictive tools, thereby expanding the chemical and structural diversity of datasets. Moreover, publicly available negative data can prevent other researchers from repeating unsuccessful designs, thus helping to save resources and time. The importance of publishing negative data has been acknowledged within the scientific community.[23,24] While many reputable journals are reluctant to publish negative results, several others explicitly encourage the dissemination thereof, including *ACS Omega*, *PLOS One*, and the *International Journal of Negative Results*. The primary objective of these journals is to reduce publication bias, enhance scientific transparency, and provide a venue for studies that contribute valuable insights despite yielding non-confirmatory or null findings. Other recommended means of publishing negative data include pre-print servers (e.g., *bioRxiv*), which enable rapid dissemination in a citable form.[25] Methodological rigor and reproducibility are the primary criteria for publishing negative data, as the absence of an effect must be unambiguously distinguished from experimental limitations.

Central Role of Raw Data

Scientific publications typically report the experimentally measured biophysical and biochemical parameters (primary data, such as T_m, k_{cat} or K_M), along with the conditions under which the measurements were performed, such as temperature, pH, and other relevant conditions. Collectively, this contextual information constitutes the metadata, encompassing experimental conditions, data processing workflows, and the statistical characterization of the processed data. However, the final values of processed data (such as estimated parameters, errors, and correlations between parameters) are sensitively dependent on the method of data processing and the unprocessed experimental data itself (so-called raw data).[26] In this context, the difference between raw and primary data is that raw data are the unprocessed experimental readouts recorded by instruments, whereas primary data are the processed and model-derived biophysical or biochemical quantities extracted from those readouts. Without reporting sufficient raw data, primary data, and metadata, the published biophysical and biochemical parameters are not reproducible. Moreover, the quality and scientific value of enzymatic data are determined primarily at the level of raw data acquisition. Reproducibility depends on well-controlled experimental conditions and consistent performance across independent experiments, ensuring that observed effects reflect true enzymatic behavior rather than technical variability.[27,28] The selectivity of the analytical method is essential for unambiguously attributing the measured signal to the intended reaction, thereby minimizing interference from side reactions, assay components, or matrix effects. Sensitivity and error analysis of the assays define the usable range and signal-to-noise ratio, directly affecting the accuracy and precision of kinetic or thermodynamic parameters.[29,30] Finally, the quality and proper characterization of enzymes, substrates, cofactors, and other reagents are critical, as impurities, instability, or concentration errors directly propagate into the raw data and ultimately limit the reliability and reproducibility of the derived results.[29,30] Similarly to the Protein Data Bank, we encourage authors of scientific papers, database providers, and dataset creators to provide the raw data in machine-readable format, in addition to their primary data, as well as structured and comprehensive metadata.

Importance of Standards in Experimental Data

The utility of the Protein Data Bank was fueled by imposing protocols to which the deposited data must adhere and iteratively improving them. In enzymology and biocatalysis, no standard protocols are widely established, although sets of recommendations exist. Examples include the Standards for Reporting Enzymology Data (STRENDA) guidelines,[31,32] the STRENDA Biocatalysis guidelines,[33,34] or the protein purification guidelines from the Protein Production and Purification Partnership in Europe network (P4EU) consortium.[20]

FAIR Principles

In biochemistry, a wider implementation of the best practices in data management is much needed,[35] namely by publishing data according to the FAIR (Findable, Accessible, Interoperable, and Reusable) Data Principles.[36,37] What does this mean? *Findable*: data should be easy to locate using search engines and other discovery tools. This includes using persistent identifiers (such as Digital Object Identifier, DOI) and rich metadata that is both human and machine-readable. *Accessible*: once data are found, there should be a clear process for accessing them. This means having clear protocols for requesting access and understanding who can access the data. *Interoperable*: data should be structured and described using shared standards to enable the integration

https://doi.org/10.1021/acscatal.5c07904
ACS Catal. 2026, 16, 8858−8868

of information from independent, non-cooperating resources with minimal effort. This ensures that data can be combined, exchanged, and automatically processed by independent communities and across diverse systems and platforms. This often involves using standard formats and ontologies. *Reusable*: data should be well annotated and have clear usage licenses so that they can be easily reused in different contexts. In essence, the FAIR principles promote the idea that research data should be treated as a valuable resource that is readily available to the broader scientific community. In practice, this translates into several important steps: (i) documenting all experiments with standard templates; (ii) saving raw and processed data in non-proprietary formats, e.g., EnzymeML, SBML, CSV, ISA-Tab; (iii) annotating them with rich metadata, e.g., STRENDA-compliant metadata (see below); (iv) depositing data into recognized repositories, e.g., STRENDA DB,[18] SABIO-RK,[19] BRENDA[6] (supported by ELIXIR; https://elixir-europe.org/), and Zenodo (https://zenodo.org/); and (v) using open standards so both humans and machines can reuse the knowledge. There are collective initiatives formed by research, industry, archives, and policy-making entities and communities that aim to implement the FAIR data principles, such as GO FAIR (https://www.go-fair.org/) and the INCF network (https://www.incf.org/).

STRENDA Guidelines

The STRENDA (https://www.beilstein-institut.de/en/projects/strenda/) guidelines[31−33] provide a set of standardized recommendations for reporting enzyme kinetics data, aiming to improve transparency, reproducibility, and data reusability in enzymology and biocatalysis. These guidelines are divided into two main levels of information, which describe the Assay Conditions (Level 1A)[38] and the Enzyme Activity Data (Level 1B),[39] respectively, and can be obtained free of charge from the respective website (https://www.beilstein-institut.de/en/projects/strenda/guidelines/). They specify the minimum required information that should be included when publishing enzyme data, such as experimental conditions, enzyme information, substrate and product details, kinetic parameters, methods, and data analysis. To maximize reusability, we encourage deposition of full time-course data (measured signal evolution for substrate or product vs. time), not only the derived parameters or initial rates. Analyses should, where feasible, employ global fitting across multiple conditions (e.g., substrate concentrations, pH, and temperature) to a shared mechanistic model. The STRENDA guidelines are designed for rigorously reporting kinetic and mechanistic enzymology data. In comparison, the STRENDA Biocatalysis guidelines[33,34] are tailored specifically to biocatalysis and applied enzyme research, and they emphasize process-relevant performance metrics, reaction outcomes, and scalability under industrially relevant conditions. We recommend reporting parameter correlations and profile-likelihood confidence intervals to address identifiability, alongside residual diagnostics and information-criterion comparisons for competing rate-law hypotheses (e.g., Michaelis-Menten). Moreover, each data entry should record the assay readout, units, replicates, uncertainty model (i.e., standard deviations or confidence intervals), the limit of detection of instruments, and the calibration procedure. This enables unambiguous reuse by modelers and meta-analysts.

The EnzymeML project[37] has been designed to fully adhere to and promote the FAIR principles and the STRENDA guidelines, within the best practices in managing enzymology data. EnzymeML is an open, XML-based, machine-readable format for documenting and exchanging biochemical and biocatalytic experimental data. EnzymeML is a "container" that holds catalytic activity and kinetic measurements, experimental conditions, and metadata required by the STRENDA and STRENDA Biocatalysis guidelines, and enables the seamless transfer of data between laboratory notebooks, databases, and modeling tools.

Resources of Unexplored Experimental Data

Community's Hidden Resources. Laboratories focused on enzyme research routinely generate significant amounts of data on enzyme characterization. Most such data are left unpublished due to negative or inconclusive results. By this, we refer to failed designs that showed poorer properties than expected and were deemed unfit for publication. In other cases, some data are never published because their quantity seems insufficient, the project was terminated, or data collection protocols were poorly recorded, among many other possible reasons. Globally, highly valuable data remain unused, archived in laboratory notebooks, worksheets, and internal databases, and are never released to the community.

Pan-European COZYME Community. COmputationally assisted design of enZYMEs (COZYME, https://cozyme.eu/) is a Pan-European collaborative network focused on the computational design of enzymes. This network, funded under the COST Action, started in 2022 and consists of three working groups with the following aims: (1) computational optimization of global enzyme properties, (2) computational optimization of catalytic properties, and (3) experimental evaluation and characterization of enzyme designs. The mission of COZYME is to advance the field of enzyme engineering by developing and utilizing cutting-edge computational tools. It brings together researchers and industry stakeholders to collaboratively develop and implement state-of-the-art methods for computational enzyme design and optimization at an industrial scale. It also includes the advancement of experimental approaches to test and validate computational predictions in the laboratory, as well as the training of young researchers. This community was the ideal target for our experiment, which involved collecting experimental data and is described in the next section.

Acquisition of Experimental Data from the COZYME Community. We have previously developed two comprehensive, manually curated databases for experimental data: FireProt^{DB} (https://loschmidt.chemi.muni.cz/fireprotdb/)[40] for protein stability data, and SoluProtMut^{DB} (https://loschmidt.chemi.muni.cz/soluprotmutdb/)[41] for protein solubility data. These databases aim to provide the community with organized and curated information. By addressing numerous drawbacks found in existing databases, they enable the development of next-generation high-accuracy computational tools for predicting mutational effects on protein stability and solubility. We have previously launched campaigns to collect experimental data from the community for inclusion in FireProt^{DB}, but we have faced limited success.

Recently, we approached the COZYME community with a request to share high-quality positive and negative experimental data with us. Some members expressed concerns. Understandably, they were apprehensive about freely providing data, which they had acquired with extensive funds, manpower, and time, to publicly accessible databases. However, the general response from the specialized enzyme- and mutagenesis-focused COZYME community was very positive, revealing a strong willingness to share their stability, solubility, and activity data

https://doi.org/10.1021/acscatal.5c07904
ACS Catal. 2026, 16, 8858−8868

(specific or total), both published and unpublished. We initially presented our case and called for data collection at the COZYME meeting in 2024. We created an online document where members could freely express their interest in participating and indicate the type and amount of data they could provide (16 research groups responded). Because our initial interest was to expand FireProt^{DB} and SoluProtMut^{DB}, we created submission forms specifically for stability and solubility data (Supplementary Forms 1 and 2, respectively), which the contributing members could easily use to unambiguously report their data. To further improve participation, we also allowed submissions in a user-customized format, given that the mandatory data were present. Using our previous experience with publishing experimental data in our databases, we specified the minimum mandatory (points 1–4) and optional (point 5) information:

1. Authorship: author names, affiliation, and publication information.
2. Reference protein of interest: protein name, amino acid or, preferably, nucleotide sequence, etc.
3. Experimental conditions: expression system, experimental method (i.e., the experimental technique used to obtain the respective raw data), physical quantity, units of the measured properties, etc.
4. Measured data points: variant name, mutations in sequence, and measured parameters.
5. Optional fields were available for other typically reported information, such as additional details or alternative parameters (e.g., organism, nucleic-acid sequence, PDB ID, UniProt ID, assay type, scan rate, Tm, ΔG, ΔΔG, Cp, half-life, total concentration, soluble concentration, soluble fraction, deep mutational scanning counts and enrichment scores, etc.), but we decided to keep the overall complexity of the forms as low as possible.

After we distributed the forms, we collected data for eight months. During this period, we conducted quality control. In case of ambiguity, we curated the submission in collaboration with the contributors via e-mail communication or video calls. The remaining steps were the deposition of the data in our public databases, FireProt^{DB} and SoluProtMut^{DB}, and Zenodo. The complete process of the data collection is illustrated in Figure 1A.

In total, data from 11 laboratories were collected, which provided 536 data points (Figure 1B; Supplementary Table 1 and Supplementary Data 1). Here, we define a data point as the outcome of a single independent experiment on a specific protein variant, which provides one single quantitative value (e.g., melting temperature T_m) or a set of inter-dependent quantities (e.g., T_m and the Gibbs free energy difference ΔG). All the data received concerned enzymes. Of these, 34 solubility data points were not usable (6% of all the data), as they contained only stained SDS-PAGE gels without an exact quantification of the soluble fractions and a detailed protocol description, which are necessary for accurate and reliable data reporting. The remaining data referred to quantified information on solubility (52%), stability (29%), and total or specific activity (13%). The stability data have been deposited to FireProt^{DB} and the solubility data will soon be deposited to SoluProtMut^{DB}. Due to our initial goals, the forms were not prepared for reporting activity data. Nonetheless, some of our contributors considered this a good opportunity to report their unpublished data. These, and all the curated received data, have been compiled and deposited in Zenodo.[42]

From the usable data, 13% contained information on wild-type proteins, 22% on single-point mutants, and the majority (65%) referred to multipoint mutants or homologues mined from enzyme-mining campaigns. Interestingly, the stability data included substantially more annotations for wild-type proteins than the solubility data, as well as a greater number of repeated measurements under varying experimental conditions (different temperatures and pH values). The variant type distributions revealed different mutational landscapes in the solubility and stability datasets. Regarding the global diversity of unique proteins, 24 of them were wild-type (5%), 228 were mutants (51%), and the remaining 199 (44%) were mined homologues. The reported enzymes span four enzyme classes (EC1−EC4; Supplementary Table 1), and the dataset is well balanced with respect to mutational effects, encompassing both positive and negative values. Data contributors were primarily located across Europe, with some representation extending into Türkiye (Figure 1B).

Lessons Learned and Recommendations for Future Efforts

Call to Action. In the data collection campaign described here, we obtained a large set of high-quality experimental data on the solubility, stability, and activity of various enzymes and their mutants (Supplementary Table 1 and Supplementary Data 1). Importantly, the datasets included a balanced ratio of positive and negative results under their respective applied experimental conditions. This would be impossible to obtain from regular literature surveys, which tend to report predominantly positive results. For permanent changes in this respect, a global shift in the current culture is needed. The data collected here represents a promising starting point, but future efforts should include more diverse communities to enhance the generalizability of our approach. Scientific journals and publishers must promote FAIR principles for data sharing. While journals often require publishing data in a supplement or as individual datasets, such data are scattered and difficult to find. Similarly to the deposition of experimental protein structures in the Protein Data Bank, the deposition of biochemical, biophysical, catalytic, and kinetic data in public databases must become a standard requirement for publication. We would like to see the process undertaken by the Protein Data Bank during the last decades of the past century, where deposition requirements inspired the community to actually share their data, as we would like to see mirrored in the biocatalysis community. In the case of the Protein Data Bank, the International Union of Crystallography first published their "Policy on Publication and the Deposition of Data from Crystallographic Studies of Biological Macromolecules",[43] which was shortly after adopted by the NIH[44] and the community in the form of journal enforcement (reviewed by Berman et al.[45]).

Enabling Data Sharing at Scale. The first obstacle that scientists encounter in sharing their data is the lack of suitable deposition venues for the task. Positive results often find their way to the wider scientific community, scattered across multiple articles, which can subsequently be compiled into databases. However, negative data are frequently buried in laboratory notebooks and left unpublished, making them inaccessible to the broader community. These high-quality laboratory data are extremely valuable for training and testing novel predictive tools. Therefore, databases that explicitly include negative data (supported by appropriate repositories, metadata standards, and incentives as listed below) are key for making this otherwise hidden treasure usable and enabling sustainable, large-scale community data collection. Our experience shows that even a simple standardized form (see above and Supplementary Forms 1 and 2), and a correct incentive

https://doi.org/10.1021/acscatal.5c07904
ACS Catal. 2026, 16, 8858−8868

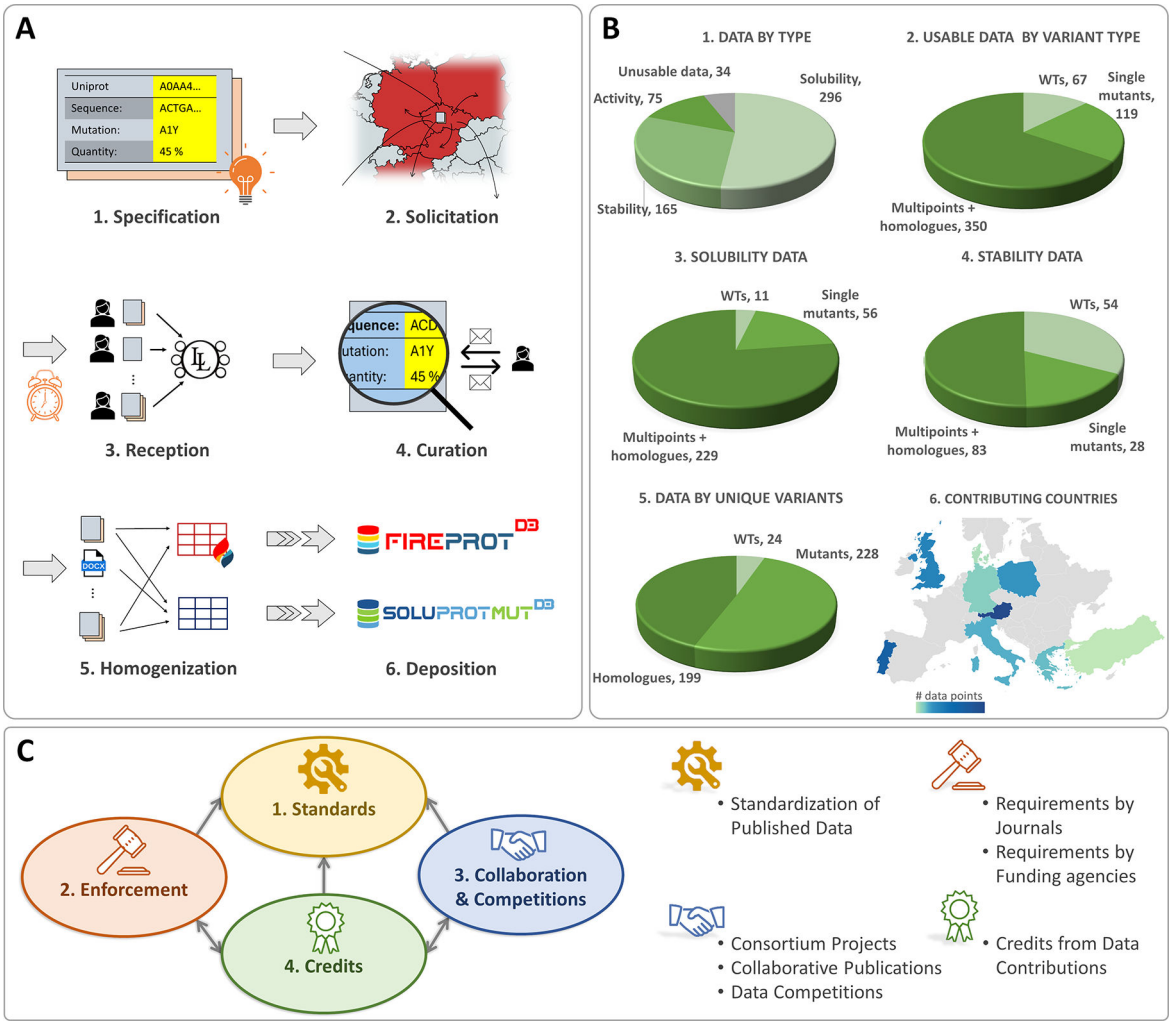


Figure 1. Community efforts to identify, curate, and deposit experimental data on mutational effects on enzyme properties. (A) Workflow of data collection from the COZYME network. This workflow depicts the main steps of the process starting with: (1) specifying the required annotations, (2) distributing the forms to the community, (3) receiving data submissions, (4) quality control and data curation, (5) homogenization of the data formats per data type, and (6) deposition of the collected data into the public databases. (B) Statistical analysis on the collected data, with its distribution by type (enzyme property), by variant (wild-type, mutants, and mined homologues), and the contributors' locations in Europe (Austria, Denmark, Germany, Greece, Italy, Poland, Portugal, and the UK) and Western Asia (Türkiye). (C) Possible future incentives to promote data sharing and reciprocal contributions within the scientific community. These incentives can take several forms: (1) establishing standards for data formats; (2) enforcement by journals and funding agencies to ensure data deposition in a standardized format; (3) community-driven initiatives, joint publications, and competitions; and (4) implementation of a credit system for shared experimental data. The arrows in the workflow illustrate their interdependencies, with the direction indicating which incentive promotes another.

can encourage members of a small community to share their experimental data. However, as emphasized by our contributors, simplifying protocols and easing data-sharing requirements would facilitate the process and promote broader participation in public repositories. Therefore, reaching an appropriate balance between minimal requirements and streamlined procedures is essential for the successful and widespread adoption of data sharing.

Toward Better Predictive Tools. Increasing the accuracy of next-generation computational tools for enzyme engineering requires training predictive models on more extensive, diverse, balanced, and high-quality experimental data. Learning from the success of AlphaFold, this can be achieved with aggregated community-sourced data. In an era where machine learning is gaining momentum, the demand for large, reliable, diverse, and balanced datasets is higher than ever. Only with such data will accurate predictions be possible across diverse enzyme families. However, reaching this point requires profound changes in how data are made available to and by the community. The COZYME community is dedicated to the development and validation of computational methods for designing enzymes. Hence, members are well aware of the need to improve state-of-the-art methods, and this has had a positive influence on the amount of data gathered in our current campaign (Supplementary Table 1 and Supplementary Data 1).

Incentives to Increase Experimental Data Sharing. An increase in the amount and quality of community-sourced positive and negative experimental data is essential. Given the collaborative nature of its objectives, the COZYME community is inclined toward data sharing. However, extending such data collection initiatives to broader scientific communities may present challenges. We believe that a shift in the current situation towards that goal can be attained through incentives and policy changes. Here, we

https://doi.org/10.1021/acscatal.5c07904
ACS Catal. 2026, 16, 8858−8868

propose several approaches that could promote the widespread sharing of data, which are related and interconnected (Figure 1C):

1. Standardization of published data: The data from biocatalytic and enzymology experiments should be reported in a standardized format for better processing using computers (e.g., EnzymeML). An XML-based data exchange format, designed to support the comprehensive documentation of enzymatic experiments and their results, should become the default means to publish experimental data. This data format observes the FAIR principles and can store detailed information, including the reaction conditions, time-course data for substrate and product concentrations, kinetic models, and the resulting estimated kinetic parameters. Furthermore, it can be easily parsed by routine programs. The standardization of published data (e.g., applying STRENDA or STRENDA Biocatalysis guidelines) is a crucial working tool and the first step towards widespread sharing of valuable experimental data.
2. Policies of peer-reviewed journals: Peer-reviewed scientific journals should require the deposition of experimental data into standardized, FAIR-compliant databases upon manuscript submission or acceptance. Editorial boards can help by supporting policies that make full data deposition a norm rather than an exception. For instance, kinetic studies shall deposit raw time courses and analysis scripts, as parameter-only deposition is insufficient. Authors should provide provisional accession numbers or DOIs for each data type, and a Data Availability Statement enumerating repositories, formats, and licenses. Since the introduction of the STRENDA Guidelines, sustained community efforts have already resulted in more than 60 biochemistry journals recommending compliance with the guidelines in their author instructions, and more than 30 journals mandating deposition of enzymatic data in the STRENDA Database (https://www.beilstein-institut.de/en/projects/strenda/). Negative outcomes are highly encouraged components of each dataset. The requirement of raw data to facilitate alternative data processing will increase reusability and prevent fraud. An example of a successful enforcement is the requirement for deposition of experimental protein structures in the Protein Data Bank as a prerequisite for publication, introduced following a community petition in 1989.[43] Platforms such as Zenodo enable this approach and can host several tens of gigabytes of experimental data. In Table 1 the preferred formats and deposition targets are presented for different types of data.
3. Policies of funding agencies: Funding bodies can promote data sharing by including it as a criterion in grant evaluations and providing necessary funding for data curation, deposition, and maintenance. Publicly funded projects, such as COZYME, can implement analogous actions for data collection, curation, and sharing. Such actions should be supported and encouraged by grant providers.
4. Collaborative publications: Consortia are a privileged means to enable the publication of joint, community-driven, peer-reviewed articles that aggregate biochemical and mutational data across multiple research groups. Such crowdsourced efforts can be organized similarly to consortium-centered studies, where contributors are credited through co-authorship or acknowledgement. A successful example is the P4EU-led community validation of Protein Repair One Stop Shop.[46]
5. Community platforms and competitions: The launching of open data challenges, competitions, or benchmarks using shared datasets can promote engagement from both experimentalists and developers of predictive tools. Examples include the ongoing protein binder design proposal by Adaptyv Bio (https://www.adaptyvbio.com/)[47] and the protein structure predictions by the Critical Assessment of Structure Prediction (CASP).[48,49] These are ways of generating highly homogenous and balanced data through community efforts, e.g., Adaptyv Bio expressed 400 designs by 130 groups in 2024 (1000 designs by 650 teams in 2025), and tested them under standardized conditions. This approach can both spotlight contributors and stimulate methodological innovations.
6. Recognition systems: A formal credit system for data sharing, analogous to Publons for peer review and now integrated into Web of Science, could incentivize researchers to deposit well-curated datasets in public repositories. Contributors could earn citations, ORCID-linked metrics, or badges reflecting their contribution to Open Science. One example is the Data Optimization Model Evaluation (DOME) registry (https://registry.dome-ml.org/), which highlights top contributors who publish machine learning

Table 1. Data Types, Preferred Formats, and Deposition Targets[a]

data type	preferred formats	primary repository	minimal data[b]
kinetics	**EnzymeML**; SBML/PEtab; CSV for raw matrices; analysis scripts	**STRENDA DB**; **SABIO RK**; BRENDA	rate law; initial concentrations (of enzyme, substrate(s), cofactors, and modifiers); kinetic model selection; confidence intervals; raw time courses; analysis code and computational environment (software, version, dependencies)
stability	**EnzymeML**; CSV; analysis notebook	**FireProt**^{DB}, ProTherm	raw measurements; stability model; calibration; assay modality; replicates; uncertainty
solubility and expressibility	**CSV** with defined readout; ontology terms	**SoluProtMut**^{DB}	experimental readout (absorbance, fluorescence, activity class, etc.); thresholds; controlled vocabulary category; host/induction

[a]Here, we list the preferred formats and repositories for the different types of data. Zenodo is recommended as an alternative and general repository for metadata and raw data. [b]Minimal data refers to the information required to interpret, reproduce, and reuse the data; additional domain-specific metadata may be required depending on the experimental context.

https://doi.org/10.1021/acscatal.5c07904
ACS Catal. 2026, 16, 8858−8868

methods that follow the standardized reporting guidelines recommended by DOME.[50,51] Datasets should be assigned DOIs, as done by repositories such as the Protein Data Bank or Zenodo, and linked to contributors via ORCID under permissive licenses (e.g., CC BY 4.0) to ensure recognition and reuse. STRENDA DB is a good example of a recognition system already in place, since: (1) it assigns a citable DOI to each deposited dataset, (2) it certifies the compliance of the data with the minimum reporting requirements, and (3) more than 30 journals require authors to deposit their data into the STRENDA DB as part of the manuscript submission process (https://www.beilstein-institut.de/en/projects/strenda/).[18] Journals and funders should recognize dataset DOIs as first-class research outputs and acknowledge reviewer credits for dataset curation.

7. In practice, individual laboratories may encounter difficulties in adopting the standards proposed herein, primarily due to time and financial constraints, as well as the need to integrate raw data collection into their routine workflows. These challenges can be mitigated through the development and widespread use of automated data capture tools and standardized submission formats.

CONCLUSIONS

To achieve further breakthroughs in the fields of biocatalysis, protein engineering, metabolic engineering and synthetic biology, enhanced reporting and standardization of high-quality experimental data are essential to improve data availability, comparability, and reuse. Our experience with the specialized Pan-European COZYME community network enabled us to gather a significant amount of data on enzyme stability, solubility, and activity, and it taught us several valuable lessons. For effective long-term improvements, multiple changes must be in place. Here, we propose several actions that can promote such a shift in the scientific culture. Some examples include the default proceedings for publishing data in scientific papers, along with the standardization and FAIRness of the published data. Sustainability requires governance beyond a single project cycle and funding models that combine modest community membership fees with competitive grants focused on interoperability. An example of this model is ELIXIR (https://elixir-europe.org/), an intergovernmental organization that brings together life science resources from across Europe, coordinating them so that they form a single infrastructure. ELIXIR is partly funded through state member contributions (often channeled through member organizations) and also supported by European and international grants. The successful implementation of some of the changes proposed here will permanently benefit the entire scientific community and biotechnology industries.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acscatal.5c07904.

Supplementary form 1 – form for stability data (XLSX)

Supplementary form 2 – form for solubility data (XLSX)

Supplementary data 1 – complete set of COZYME data collected (XLSX)

Supplementary table 1 – summary of enzymes from the COZYME data collection (XLSX)

AUTHOR INFORMATION

Corresponding Authors

Jiri Damborsky – *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Kotlářská 267/2, Brno 611 37, Czech Republic; International Clinical Research Center, St. Anne's University Hospital Brno, Pekařská 53, Brno 602 00, Czech Republic;* Email: jiri@chemi.muni.cz

Stanislav Mazurenko – *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Kotlářská 267/2, Brno 611 37, Czech Republic; International Clinical Research Center, St. Anne's University Hospital Brno, Pekařská 53, Brno 602 00, Czech Republic;* orcid.org/0000-0003-3659-4819; Email: mazurenko@mail.muni.cz

David Bednar – *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Kotlářská 267/2, Brno 611 37, Czech Republic; International Clinical Research Center, St. Anne's University Hospital Brno, Pekařská 53, Brno 602 00, Czech Republic;* orcid.org/0000-0002-6803-0340; Email: 222755@mail.muni.cz

Authors

Sérgio M. Marques – *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Kotlářská 267/2, Brno 611 37, Czech Republic; International Clinical Research Center, St. Anne's University Hospital Brno, Pekařská 53, Brno 602 00, Czech Republic;* orcid.org/0000-0002-6281-7505

Joan Planas-Iglesias – *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Kotlářská 267/2, Brno 611 37, Czech Republic; International Clinical Research Center, St. Anne's University Hospital Brno, Pekařská 53, Brno 602 00, Czech Republic;* orcid.org/0000-0002-6279-2483

Jan Velecký – *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Kotlářská 267/2, Brno 611 37, Czech Republic; International Clinical Research Center, St. Anne's University Hospital Brno, Pekařská 53, Brno 602 00, Czech Republic*

Milos Musil – *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Kotlářská 267/2, Brno 611 37, Czech Republic; International Clinical Research Center, St. Anne's University Hospital Brno, Pekařská 53, Brno 602 00, Czech Republic; Department of Information Systems, Faculty of Information Technology, Brno University of Technology, Božetěchova 2, Brno 612 00, Czech Republic*

Yasuhisa Asano – *Biotechnology Research Center and Department of Biotechnology, Toyama Prefectural University, 5180 Kurokawa, Imizu, Toyama 939-0398, Japan;* orcid.org/0000-0003-3645-3952

Tomasz Borowski – *Jerzy Haber Institute of Catalysis and Surface Chemistry, Polish Academy of*

https://doi.org/10.1021/acscatal.5c07904
ACS Catal. 2026, 16, 8858−8868

Sciences, Niezapominajek 8, Krakow 30-239, Poland; orcid.org/0000-0002-3450-3576

Vânia Brissos – *Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Av. da República, Oeiras 2780-157, Portugal*

Marco Cespugli – *Innophore GmbH, Am Eisernen Tor 3, Graz 8010, Austria*

Koar Chorozian – *School of Chemical Engineering, National Technical University of Athens, 9 Iroon Polytechniou Str, Zografou GR-15772, Greece*

Mohammad Dadashipour – *School of Health and Life Sciences, Teesside University, Middlesbrough TS1 3BA, U.K.; National Horizons Centre, Teesside University, Darlington DL1 1HG, U.K.*

Elif Erdem – *Department of Chemical and Biochemical Engineering, Technical University of Denmark, Kgs., Lyngby 2800, Denmark;* orcid.org/0000-0003-1411-9793

Erica Elisa Ferrandi – *Istituto di Scienze e Tecnologie Chimiche "Giulio Natta" (SCITEC) — CNR, Via Mario Bianco 9, Milano 20131, Italy;* orcid.org/0000-0002-3390-9638

Konstantinos Grigorakis – *School of Chemical Engineering, National Technical University of Athens, 9 Iroon Polytechniou Str, Zografou GR-15772, Greece*

Anna Kluza – *Jerzy Haber Institute of Catalysis and Surface Chemistry, Polish Academy of Sciences, Niezapominajek 8, Krakow 30-239, Poland*

Janina Lawniczek – *Microbial Biotechnology, Faculty of Biology and Biotechnology, Ruhr University Bochum, Universitätsstraße 150, Bochum 44780, Germany*

Konstantinos Makryniotis – *School of Chemical Engineering, National Technical University of Athens, 9 Iroon Polytechniou Str, Zografou GR-15772, Greece*

Daniela Monti – *Istituto di Scienze e Tecnologie Chimiche "Giulio Natta" (SCITEC) — CNR, Via Mario Bianco 9, Milano 20131, Italy;* orcid.org/0000-0002-3399-7973

Bettina Nestl – *Innophore GmbH, Am Eisernen Tor 3, Graz 8010, Austria;* orcid.org/0000-0003-1282-195X

Anna C. Ngo – *Microbial Biotechnology, Faculty of Biology and Biotechnology, Ruhr University Bochum, Universitätsstraße 150, Bochum 44780, Germany*

Efstratios Nikolaivits – *School of Chemical Engineering, National Technical University of Athens, 9 Iroon Polytechniou Str, Zografou GR-15772, Greece;* orcid.org/0000-0002-8022-9272

Stefania Patti – *Istituto di Scienze e Tecnologie Chimiche "Giulio Natta" (SCITEC) — CNR, Via Mario Bianco 9, Milano 20131, Italy*

Christina Pentari – *School of Chemical Engineering, National Technical University of Athens, 9 Iroon Polytechniou Str, Zografou GR-15772, Greece;* orcid.org/0000-0003-3478-8812

Carolina F. Rodrigues – *Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Av. da República, Oeiras 2780-157, Portugal*

Tobias Schopper – *Innophore GmbH, Am Eisernen Tor 3, Graz 8010, Austria*

Karolina Seweryn-Ożóg – *Jerzy Haber Institute of Catalysis and Surface Chemistry, Polish Academy of Sciences, Niezapominajek 8, Krakow 30-239, Poland*

Maciej Szaleniec – *Jerzy Haber Institute of Catalysis and Surface Chemistry, Polish Academy of Sciences, Niezapominajek 8, Krakow 30-239, Poland;* orcid.org/0000-0002-7650-9263

André Taborda – *Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Av. da República, Oeiras 2780-157, Portugal*

Mateusz Tataruch – *Jerzy Haber Institute of Catalysis and Surface Chemistry, Polish Academy of Sciences, Niezapominajek 8, Krakow 30-239, Poland*

Dirk Tischler – *Microbial Biotechnology, Faculty of Biology and Biotechnology, Ruhr University Bochum, Universitätsstraße 150, Bochum 44780, Germany;* orcid.org/0000-0002-6288-2403

Evangelos Topakas – *School of Chemical Engineering, National Technical University of Athens, 9 Iroon Polytechniou Str, Zografou GR-15772, Greece;* orcid.org/0000-0003-0078-5904

Jingyu Wang – *Department of Chemical and Biochemical Engineering, Technical University of Denmark, Kgs., Lyngby 2800, Denmark*

Patrycja Wójcik – *Jerzy Haber Institute of Catalysis and Surface Chemistry, Polish Academy of Sciences, Niezapominajek 8, Krakow 30-239, Poland*

Agnieszka M. Wojtkiewicz – *Jerzy Haber Institute of Catalysis and Surface Chemistry, Polish Academy of Sciences, Niezapominajek 8, Krakow 30-239, Poland*

John M. Woodley – *Department of Chemical and Biochemical Engineering, Technical University of Denmark, Kgs., Lyngby 2800, Denmark;* orcid.org/0000-0002-7976-2483

Olga Zastawny – *Jerzy Haber Institute of Catalysis and Surface Chemistry, Polish Academy of Sciences, Niezapominajek 8, Krakow 30-239, Poland*

Lígia O. Martins – *Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Av. da República, Oeiras 2780-157, Portugal*

Marco Fraaije – *Molecular Enzymology group, Molecular Enzymology group, Nijenborgh 4, Groningen 9747AG, The Netherlands;* orcid.org/0000-0001-6346-5014

Jürgen Pleiss – *Institute of Biochemistry, University of Stuttgart, Allmandring 31, Stuttgart 70569, Germany;* orcid.org/0000-0003-1045-8202

Santiago Schnell – *Department of Mathematics, Dartmouth College, and Department of Biochemistry & Cell Biology, and Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire 03755, USA;* orcid.org/0000-0002-9477-3914

Complete contact information is available at: https://pubs.acs.org/doi/10.1021/acscatal.5c07904

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to express their thanks to the COST Action CA21162 (COZYME) funded by the European Union's Cooperation in Science and Technology. This project has also received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements

No. 857560 (CETOCOEN) and Horizon Europe Framework programme No. 101136607 (CLARA). Computational resources were provided by the e-INFRA CZ, ELIXIR-CZ, and RECETOX RI projects (90254, LM2023055, and LM2023069), supported by the Ministry of Education, Youth and Sports of the Czech Republic. This publication reflects only the author's view, and the European Commission is not responsible for any use that may be made of the information it contains.

GLOSSARY

BRENDA	BRaunschweig ENzyme Database with biochemical and molecular information on enzymes.
CASP	Critical Assessment of Structure Prediction, is a biennial, worldwide competition that provides a rigorous, blind test of computational protein structure modeling methods.
CC BY 4.0	is a license that allows others to copy, distribute, remix, and build upon a work, even for commercial purposes, as long as they give appropriate credit to the original creator and indicate if changes were made.
COST	European Cooperation in Science and Technology, is an EU-funded initiative that supports the creation of interdisciplinary research networks across Europe and beyond.
COZYME	COmputationally assisted design of enZYMEs, is a Pan-European collaborative action funded by COST that aims to develop and improve computational tools for enzyme design and engineering.
CSV	Comma-Separated Values, is a simple, text-based format for tabular data, where values are separated by commas and rows are separated by newlines.
DOI	Digital Object Identifier, is a unique, permanent, and persistent alphanumeric string assigned to digital resources, such as scientific articles, datasets, and books.
DOME	Data Optimization Model Evaluation, is a centralized, structured database that serves as a repository for information on published machine learning studies, particularly in biology.
EnzymeML	is an Extensible Markup Language (XML)-based data exchange format that supports the comprehensive documentation of biocatalytic data.
FAIR	Findable, Accessible, Interoperable, and Reusable, referring to scientific data.
FireProtDB	is a comprehensive, manually curated database of protein stability data.
INCF	The International Neuroinformatics Coordinating Facility network, is a global collaborative forum of researchers, institutions, companies, and publishers dedicated to advancing neuroinformatics.
ISA-Tab	Investigation, Study, Assay, is a structured, multi-file format used to describe complex experimental metadata.
P4EU	Protein Production and Purification Partnership in Europe network.
SABIO-RK	System for the Analysis of Biochemical Pathways − Reaction Kinetics, is a comprehensive, manually curated public database for biochemical reactions and their kinetics.
SBML	Systems Biology Markup Language, is an XML-based file format designed for representing and exchanging computational models of biological processes.
SI	International System of Units, is the modern form of the metric system.
SoluProtMutDB	is a comprehensive, manually curated database of protein solubility data.
STRENDA	Standards for Reporting Enzymology Data, is a set of guidelines to ensure enzyme activity and kinetic data reported in scientific publications are complete, accurate, and reproducible.
STRENDA DB	is a database that validates and stores enzyme function data based on the STRENDA Guidelines.
UniProt	Universal Protein Resource database of protein sequence and functional information.
Zenodo	is a free, open-access repository that allows researchers to share and preserve various research outputs, including datasets, software, and publications, in any file format.

REFERENCES

(1) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589.

(2) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.

(3) Kurisu, G. Fifty Years of Protein Data Bank in the Journal of Biochemistry. *J. Biochem.* **2022**, *171* (1), 3–11.

(4) Berman, H. M.; Burley, S. K. Protein Data Bank (PDB): Fifty-Three Years Young and Having a Transformative Impact on Science and Society. *Q. Rev. Biophys.* **2025**, *58*, No. e9.

(5) The UniProt ConsortiumUniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* **2025**, *53* (D1), D609–D617.

(6) Chang, A.; Jeske, L.; Ulbrich, S.; Hofmann, J.; Koblitz, J.; Schomburg, I.; Neumann-Schaal, M.; Jahn, D.; Schomburg, D. BRENDA, the ELIXIR Core Data Resource in 2021: New Developments and Updates. *Nucleic Acids Res.* **2021**, *49* (D1), D498–D508.

(7) Kulmanov, M.; Khan, M. A.; Hoehndorf, R. DeepGO: Predicting Protein Functions from Sequence and Interactions Using a Deep Ontology-Aware Classifier. *Bioinformatics* **2018**, *34* (4), 660–668.

(8) Gligorijević, V.; Renfrew, P. D.; Kosciolek, T.; Leman, J. K.; Berenberg, D.; Vatanen, T.; Chandler, C.; Taylor, B. C.; Fisk, I. M.; Vlamakis, H.; Xavier, R. J.; Knight, R.; Cho, K.; Bonneau, R. Structure-Based Protein Function Prediction Using Graph Convolutional Networks. *Nat. Commun.* **2021**, *12* (1), 3168.

(9) Boadu, F.; Lee, A.; Cheng, J. Deep Learning Methods for Protein Function Prediction. *Proteomics* **2025**, *25* (1−2), No. 2300471.

(10) Herencias, C.; Salgado-Briegas, S.; Prieto, M. A.; Nogales, J. Providing New Insights on the Biphasic Lifestyle of the Predatory Bacterium Bdellovibrio Bacteriovorus through Genome-Scale Metabolic Modeling. *PLoS Comput. Biol.* **2020**, *16* (9), No. e1007646.

(11) Nakamura, T.; Fahmi, M.; Tanaka, J.; Seki, K.; Kubota, Y.; Ito, M. Genome-Wide Analysis of Whole Human Glycoside Hydrolases by Data-Driven Analysis in Silico. *Int. J. Mol. Sci.* **2019**, *20* (24), 6290.

(12) Domenzain, I.; Sánchez, B.; Anton, M.; Kerkhoven, E. J.; Millán-Oropeza, A.; Henry, C.; Siewers, V.; Morrissey, J. P.;

https://doi.org/10.1021/acscatal.5c07904
ACS Catal. 2026, 16, 8858−8868

Sonnenschein, N.; Nielsen, J. Reconstruction of a Catalogue of Genome-Scale Metabolic Models with Enzymatic Constraints Using GECKO 20. *Nat. Commun.* **2022**, *13* (1), 3766.
(13) Gupta, A.; Agrawal, S. Machine Learning-Based Enzyme Engineering of PETase for Improved Efficiency in Plastic Degradation. *J. Emerg. Invest.* **2023**, 16.
(14) Ming, Y.; Wang, W.; Yin, R.; Zeng, M.; Tang, L.; Tang, S.; Li, M. A Review of Enzyme Design in Catalytic Stability by Artificial Intelligence. *Brief Bioinform.* **2023**, *24* (3), No. bbad065.
(15) Sumida, K. H.; Núñez-Franco, R.; Kalvet, I.; Pellock, S. J.; Wicky, B. I. M.; Milles, L. F.; Dauparas, J.; Wang, J.; Kipnis, Y.; Jameson, N.; Kang, A.; De La Cruz, J.; Sankaran, B.; Bera, A. K.; Jiménez-Osés, G.; Baker, D. Improving Protein Expression, Stability, and Function with ProteinMPNN. *J. Am. Chem. Soc.* **2024**, *146* (3), 2054–2061.
(16) Wang, Z.; Xie, D.; Wu, D.; Luo, X.; Wang, S.; Li, Y.; Yang, Y.; Li, W.; Zheng, L. Robust Enzyme Discovery and Engineering with Deep Learning Using CataPro. *Nat. Commun.* **2025**, *16* (1), No. 2736.
(17) Nestl, B. M.; Nebel, B. A.; Resch, V.; Schürmann, M.; Tischler, D. The Development and Opportunities of Predictive Biotechnology. *ChemBioChem* **2024**, *25* (13), No. e202300863.
(18) Swainston, N.; Baici, A.; Bakker, B. M.; Cornish-Bowden, A.; Fitzpatrick, P. F.; Halling, P.; Leyh, T. S.; O'Donovan, C.; Raushel, F. M.; Reschel, U.; Rohwer, J. M.; Schnell, S.; Schomburg, D.; Tipton, K. F.; Tsai, M.-D.; Westerhoff, H. V.; Wittig, U.; Wohlgemuth, R.; Kettner, C. STRENDA DB: Enabling the Validation and Sharing of Enzyme Kinetics Data. *FEBS J.* **2018**, *285* (12), 2193–2204.
(19) Wittig, U.; Rey, M.; Weidemann, A.; Kania, R.; Müller, W. SABIO-RK: An Updated Resource for Manually Curated Biochemical Reaction Kinetics. *Nucleic Acids Res.* **2018**, *46* (D1), D656–D660.
(20) Berrow, N.; de Marco, A.; Lebendiker, M.; Garcia-Alai, M.; Knauer, S. H.; Lopez-Mendez, B.; Matagne, A.; Parret, A.; Remans, K.; Uebel, S.; Raynal, B. Quality Control of Purified Proteins to Improve Data Quality and Reproducibility: Results from a Large-Scale Survey. *Eur. Biophys. J.* **2021**, *50* (3), 453–460.
(21) Krauss, U.; Jäger, V. D.; Diener, M.; Pohl, M.; Jaeger, K.-E. Catalytically-Active Inclusion Bodies—Carrier-Free Protein Immobilizates for Application in Biotechnology and Biomedicine. *J. Biotechnol.* **2017**, *258*, 136–147.
(22) Herbet, M.-E.; Leonard, J.; Santangelo, M. G.; Albaret, L. Dissimulate or Disseminate? A Survey on the Fate of Negative Results. *Learned Publ.* **2022**, *35* (1), 16–29.
(23) Echevarría, L.; Malerba, A.; Arechavala-Gomeza, V. Researcher's Perceptions on Publishing "Negative" Results and Open Access. *Nucleic Acid Ther.* **2021**, *31* (3), 185–189.
(24) Curry, S.; Mercado-Lara, E.; Arechavala-Gomeza, V.; Begley, C. G.; Bernard, C.; Bernard, R.; Bertuzzi, S.; Bhalla, N.; Bowers, D.; Brod, S.; Chambers, C.; Dougherty, M. R.; Bueso, Y. F.; Forner, S.; Freeman, A. L. J.; Haas, M.; Henderson, D. P.; Khanna, K.; Lawrence, R.; Liakath-Ali, K.; Liu, C.; Malhotra, N.; Merino, J. G.; Miguel, E.; Miles, R.; Munson, M.; Nakagawa, S.; Nobles, R.; Owango, J.; Pham, M. T.; Poe, G.; Ramirez, A. N.; Sarabipour, S.; Silverman, J. L.; Smith, L. N.; Sriramarao, P.; Sternberg, P. W.; Swamy, G. K.; Tansey, M. G.; Torres, G. E.; Turner, E. H.; Klinggraeff, L. von; Weis-Garcia, F. Ending Publication Bias: A Values-Based Approach to Surface Null and Negative Results. *PLoS Biol.* **2025**, *23* (9), No. e3003368.
(25) Nimpf, S.; Keays, D. A. Why (and How) We Should Publish Negative Data. *EMBO Rep.* **2020**, *21* (1), No. e49775.
(26) Duggleby, R. G. Experimental Designs for Estimating the Kinetic Parameters for Enzyme-Catalysed Reactions. *J. Theor. Biol.* **1979**, *81* (4), 671–684.
(27) Halling, P.; Fitzpatrick, P. F.; Raushel, F. M.; Rohwer, J.; Schnell, S.; Wittig, U.; Wohlgemuth, R.; Kettner, C. An Empirical Analysis of Enzyme Function Reporting for Experimental Reproducibility: Missing/Incomplete Information in Published Papers. *Biophys. Chem.* **2018**, *242*, 22–27.
(28) Gygli, G. On the Reproducibility of Enzyme Reactions and Kinetic Modelling. *Biol. Chem.* **2022**, *403* (8–9), 717–730.
(29) Gardossi, L.; Poulsen, P. B.; Ballesteros, A.; Hult, K.; Svedas, V. K.; Vasić-Racki, D.; Carrea, G.; Magnusson, A.; Schmid, A.; Wohlgemuth, R.; Halling, P. J. Guidelines for Reporting of Biocatalytic Reactions.European Federation of Biotechnology Section on Applied Biocatalysis *Trends Biotechnol.* **2010**, *28* (4), 171–180.
(30) de Marco, A.; Berrow, N.; Lebendiker, M.; Garcia-Alai, M.; Knauer, S. H.; Lopez-Mendez, B.; Matagne, A.; Parret, A.; Remans, K.; Uebel, S.; Raynal, B. Quality Control of Protein Reagents for the Improvement of Research Data Reproducibility. *Nat. Commun.* **2021**, *12* (1), No. 2795.
(31) Tipton, K. F.; Armstrong, R. N.; Bakker, B. M.; Bairoch, A.; Cornish-Bowden, A.; Halling, P. J.; Hofmeyr, J.-H.; Leyh, T. S.; Kettner, C.; Raushel, F. M.; Rohwer, J.; Schomburg, D.; Steinbeck, C. Standards for Reporting Enzyme Data: The STRENDA Consortium: What It Aims to Do and Why It Should Be Helpful. *Perspect. Sci.* **2014**, *1* (1), 131–137.
(32) Goldberg, R. N.; Giessmann, R. T.; Halling, P. J.; Kettner, C.; Westerhoff, H. V. Recommendations for Performing Measurements of Apparent Equilibrium Constants of Enzyme-Catalyzed Reactions and for Reporting the Results of These Measurements. *Beilstein J. Org. Chem.* **2023**, *19*, 303–316.
(33) Malzacher, S.; Meißner, D.; Range, J.; Findrik Blažević, Z.; Rosenthal, K.; Woodley, J. M.; Wohlgemuth, R.; Wied, P.; Nidetzky, B.; Giessmann, R. T.; Prakinee, K.; Chaiyen, P.; Bommarius, A. S.; Rohwer, J. M.; de Souza, R. O. M. A.; Halling, P. J.; Pleiss, J.; Kettner, C.; Rother, D. The STRENDA Biocatalysis Guidelines for Cataloguing Metadata. *Nat. Catal.* **2024**, *7* (12), 1245–1249.
(34) Meißner, D.; Stephan, S.; Range, J. *Strenda-Biocatalysis/Strenda-Biocatalysis, GitHub, 2024.* https://github.com/Strenda-biocatalysis/Strenda-biocatalysis (*accessed January 30, 2026*)
(35) Giess, T.; Pleiss, J. Chapter Two - Digitalization of Biocatalysis: Best Practices to Research Data Management. In *Methods Enzymol.*, Tischler, D., Ed.; Biocatalysis Identifying novel enzymes and applying them in cell-free and whole-cell biocatalysis; Academic Press, **2025**; Vol. *714*, pp 19–43
(36) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. j. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; de Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; 't Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3* (1), No. 160018.
(37) Pleiss, J. Standardized Data, Scalable Documentation, Sustainable Storage – EnzymeML As A Basis For FAIR Data Management In Biocatalysis. *ChemCatChem* **2021**, *13* (18), 3909–3913.
(38) Beilstein STRENDA Commission. *STRENDA Guideline Level 1A Experimental Conditions, Version 1.8, SEC.3.* https://doi.org/10.3762/strenda.18.
(39) Beilstein STRENDA Commission. *STRENDA Guideline Level 1B Experimental Results, Version 1.8, SER.3.* https://doi.org/10.3762/strenda.28.
(40) Stourac, J.; Dubrava, J.; Musil, M.; Horackova, J.; Damborsky, J.; Mazurenko, S.; Bednar, D. FireProtDB: Database of Manually Curated Protein Stability Data. *Nucleic Acids Res.* **2021**, *49* (D1), D319–D324.
(41) Velecký, J.; Hamsikova, M.; Stourac, J.; Musil, M.; Damborsky, J.; Bednar, D.; Mazurenko, S. SoluProtMutDB: A Manually Curated Database of Protein Solubility Changes upon Mutations. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 6339–6347.

https://doi.org/10.1021/acscatal.5c07904
ACS Catal. 2026, 16, 8858–8868

(42) Musil, M.; Marques, S. Data Collected through COZYME Initiative (1.1) [Data set]. *Zenodo* **2026**, https://doi.org/10.5281/zenodo.18428322.
(43) Commission on Biological Macromolecules Policy on Publication and the Deposition of Data from Crystallographic Studies of Biological Macromolecules. *Acta Cryst.* **1989**, *A45*, 658.
(44) *NIH Guide*: *PUBLIC HEALTH SERVICE POLICY RELATING TO DISTRIBUTION OF UNIQUERESEARCH RESOURCES PRODUCED WITH PHS FUNDING.* https://grants.nih.gov/grants/guide/notice-files/not92-163.html (accessed January 30, 2026).
(45) Berman, H. M.; Vallat, B.; Lawson, C. L. The Data Universe of Structural Biology. *IUCrJ* **2020**, *7* (4), 630–638.
(46) Peleg, Y.; Vincentelli, R.; Collins, B. M.; Chen, K.-E.; Livingstone, E. K.; Weeratunga, S.; Leneva, N.; Guo, Q.; Remans, K.; Perez, K.; Bjerga, G. E. K.; Larsen, Ø.; Vaněk, O.; Skořepa, O.; Jacquemin, S.; Poterszman, A.; Kjær, S.; Christodoulou, E.; Albeck, S.; Dym, O.; Ainbinder, E.; Unger, T.; Schuetz, A.; Matthes, S.; Bader, M.; de Marco, A.; Storici, P.; Semrau, M. S.; Stolt-Bergner, P.; Aigner, C.; Suppmann, S.; Goldenzweig, A.; Fleishman, S. J. Community-Wide Experimental Evaluation of the PROSS Stability-Design Method. *J. Mol. Biol.* **2021**, *433* (13), No. 166964.
(47) Cotet, T.-S.; Krawczuk, I.; Stocco, F.; Ferruz, N.; Gitter, A.; Kurumida, Y.; Machado, L. de A.; Paesani, F.; Calia, C. N.; Challacombe, C. A.; Haas, N.; Qamar, A.; Correia, B. E.; Pacesa, M.; Nickel, L.; Subr, K.; Castorina, L. V.; Campbell, M. J.; Ferragu, C.; Kidger, P.; Hallee, L.; Wood, C. W.; Stam, M. J.; Kluonis, T.; Ünal, S. M.; Belot, E.; Naka, A. Crowdsourced Protein Design: Lessons From the Adaptyv EGFR Binder Competition. Organizers, A. C. *bioRxiv* **2025**, No. 2025.04.17.648362.
(48) Tosstorff, A.; Rudolph, M. G.; Benz, J.; Kuhn, B.; Kramer, C.; Sharpe, M.; Huang, C.-Y.; Metz, A.; Hazemann, J.; Ritz, D.; Sweeney, A. M.; Gilson, M. The CASP 16 Experimental Protein-Ligand Datasets. *Proteins* **2025**, *94*, 79–85.
(49) Yuan, R.; Zhang, J.; Kryshtafovych, A.; Schaeffer, R. D.; Zhou, J.; Cong, Q.; Grishin, N. V. CASP16 Protein Monomer Structure Prediction Assessment. *Proteins Struct. Funct. Bioinform.* **2025**, *94*, 86.
(50) Walsh, I.; Fishman, D.; Garcia-Gasulla, D.; Titma, T.; Pollastri, G.; Harrow, J.; Psomopoulos, F. E.; Tosatto, S. C. E. DOME: Recommendations for Supervised Machine Learning Validation in Biology. *Nat. Methods* **2021**, *18* (10), 1122–1127.
(51) Attafi, O. A.; Clementel, D.; Kyritsis, K.; Capriotti, E.; Farrell, G.; Fragkouli, S.-C.; Castro, L. J.; Hatos, A.; Lenaerts, T.; Mazurenko, S.; Mozaffari, S.; Pradelli, F.; Ruch, P.; Savojardo, C.; Turina, P.; Zambelli, F.; Piovesan, D.; Monzon, A. M.; Psomopoulos, F.; Tosatto, S. C. E. DOME Registry: Implementing Community-Wide Recommendations for Reporting Supervised Machine Learning in Biology. *Gigascience* **2024**, *13*, No. giae094.

https://doi.org/10.1021/acscatal.5c07904
ACS Catal. 2026, 16, 8858–8868