nature portfolio

# Peer Review File

# Connecting chemical and protein sequence space to predict biocatalytic reactions

Corresponding Author: Professor Alison Narayan

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

**This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.**

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
Paton et al. describe in this interesting and well written manuscript how they attempt to connect “chemical and protein sequence space” exemplified for α-ketoglutarate-dependent non-heme iron (NHI) enzymes. Thus >100 substrates were chosen for a given enzyme and then this was experimentally explored resulting in novel biocatalytic reactions/expanded substrate scopes.
To identify the most suitable protein sequence space, they have performed what is common practice now since the Enzyme Function Initiative–Enzyme Similarity Tool was established by John Gerlt. One makes a solid in-depth sequence similarity network (SSN) analysis, which results in a set of (new) sequences encoding different enzymes, which are then studied (from synthetic genes, expressed in E. coli) to identify e.g. what substrate scope these (novel/underexplored/never studied) biocatalysts have. This results in a set of hits and a lot of experimental data, which can then be used for machine learning guided predictions, i.e. identification of positions for mutations to further control the outcome of a biocatalytic reaction (i.e. more selectivity, more stability, broader or smaller substrate scope).
Concepts like this have already been studied two decades ago (but at that time SSN etc. was not invented yet). The group of Reymond has created fingerprint profiles for esterase/lipase/epoxide hydrolase etc. using a high-throughput screening system (Wahler et al, cited below), other groups explored error-prone PCR-based mutant libraries (that was easier and cheaper 20 years ago then ordering a set of synthetic genes). The company Diversa published also 20 years ago how to explore sequence space for nitrilases (Robertson et al., cited below). More recently, the group of Pelletier made a major effort to put experimental data from numerous literature-reported reactions catalyzed by the monooxygenase P450-BM3 in a public accessible database to facilitate identification of P450 (variants) with desired features for a given substrate (Fansher et al., cited below), another group has done this for transaminases, where most notably the “chemical space” was already included in the machine learning algorithm to not only compare different enzyme sequences, but also to predict the substrate scope using machine learning (Ao et al., cited below). These are just a few of several examples from literature, where [as stated in their abstract of this manuscript] “strategies for predicting which enzyme is most likely to be compatible with a given small molecule“ have already been developed/described.

None of these papers has been cited here.

Thus, the “longstanding challenge” claimed here in the abstract has already been addressed, at least to some extent.

Another aspect, which I have missed to consider is the correct prediction of enzyme function from sequences in databases, where the Zhao group has established an interesting concept, dubbed ‘contrastive learning’ (Yu et al., cited below). Clearly, this correction of misleading or wrong annotations of protein function must be taken into account when using sequence information to figure out “chemical space” targets.
Finally, I have missed citations / discussion of two recent reviews relevant in this context (Lovelock et al., Buller et al., cited below). Both reviews also deal in part with the issue of connecting protein sequences to functon(substrate identification/prediction) especially related to the reliable prediction of enzyme properties from protein sequences (which is still challenging using the most advanced AlphaFold version 3.0).

In summary, the authors have done a great job to establish what they have described in this manuscript. For sure, this is important to other scientists. But I do not see any outstanding novel concept to address the problem of how to connect protein space to chemical space as similar or even identical strategies are documented in literature as exemplified above.

Thus, I can't recommend publication of this work in the journal Nature.

There is no need for a major revision (for submission to another journal), I only have a few points:

Ref. 23: it is stated that in 2023 only 1264 journal articles were published on biocatalysis. This must be a completely wrong number. I think this depends on the keywords used for searching (if you search for "electrochemistry", I am confident that this keyword finds most/all publications related to this), i.e. "enzyme catalysis", "enzyme", but also "biotransformation" all must be used to search in the broad area of biocatalysis. I quickly checked this in Web of Science and found 1895 journal articles (not reviews) published in 2023 and I found >20,000 papers when using also the term "enzyme". I suggest to delete Ref. 23 / this statement.

"600,000 fold improvement in yield" stated on p.3 for the research described in Ref. 30: Yield is isolated product, how can one improve this 600,000 fold as it is already a challenge to isolate (find) 0,1% of a product. Is this really yield or just rate acceleration? Please check the term "yield" before such a statement is written in the manuscript.

A side note: The authors also developed the open access interface "Catnip", which should help to use their predictive workflow. I very much appreciate this effort.
I have tried Catnip drawing a simple molecule (e.g. acetophenone) and would then expect that the first two hits would be "ketoreductase" to yield the chiral alcohol or "amine transferase" to yield the chiral amine (and maybe a list of various wildtype enzymes or some mutants or acetophenone derivatives with some substitutions at the aromatic ring). The outcome was confusing to me, as numerous highly complex (or unrelated) molecules were predicted by Catnip, but none of them related at all to the expected (most logical first enzymatic steps) suggestion of enzymes or 'chemical space'. Also, the list of enzymes in the ranking made no sense to me.

Missing examples/references:
R. Buller, S. Lutz, R. J. Kazlauskas, R. Snajdrova, J. C. Moore, U. T. Bornscheuer, Science 2023, 382, eadh8615.
S. L. Lovelock, R. Crawshaw, S. Basler, C. Levy, D. Baker, D. Hilvert, A. P. Green, The road to fully programmable protein catalysis, Nature 2022, 606, 49-58.
T. Yu, H. Cui, J. C. Li, Y. Luo, G. Jiang, H. Zhao, Enzyme function prediction using contrastive learning, Science 2023, 379, 1358-1363.
Ao, Y.F., Pei, S., Xiang, C., Menke, M.J., Shen, L., Sun, C., Dörr, M., Born, S., Höhne, M., Bornscheuer, U.T. (2023), Structure- and data-driven protein engineering of transaminases for improving activity and stereoselectivity, Angew. Chem. Int. Ed., 62, e202301660.
D. J. Fansher, J. N. Besna, A. Fendri, J. N. Pelletier, Choose Your Own Adventure: A Comprehensive Database of Reactions Catalyzed by Cytochrome P450 BM3 Variants, ACS Catal. 2024, 14, 5560-5592.
D. E. Robertson, J. A. Chaplin, G. DeSantis, M. Podar, M. Madden, E. Chi, T. Richardson, A. Milan, M. Miller, D. P. Weiner, K. Wong, J. McQuaid, B. Farwell, L. A. Preston, X. Tan, M. A. Snead, M. Keller, E. Mathur, P. L. Kretz, M. J. Burk, J. M. Short, Exploring nitrilase sequence space for enantioselective catalysis, Appl. Environ. Microbiol. 2004, 70, 2429-2436.
D. Wahler, F. Badalassi, P. Crotti, J.-L. Reymond, Enzyme fingerprints by fluorogenic and chromogenic substrate arrays, Angew. Chem. Int. Ed. 2001, 40, 4457-4460.

(Remarks on code availability)

Referee #2

(Remarks to the Author)
The manuscript by Paton et al describes an interesting study where the authors utilize EFI tools to select a broad range of 314 Fe/2OG-dependent enzymes for screening against a panel of 111 substrates. The substrates ranged from small aromatics and amino acids to larger extended and polycyclic structures to around ~500 Da in size. From this screen, a good number of hits were found with 32% of substrates being accepted by an enzyme and 38% of enzymes showing competence on a substrate, yielding 215 new reactions. Machine learning was then applied using this dataset as well as literature reactions (139 reactions) with the goal of predicting enzymes to screen for specific substrates. This approach was applied to three different test molecules, for which transformations were successfully obtained.
Overall, I enjoyed reading the manuscript and I think that the study is well-designed and well-conceived. However, I am not sure that the methods or results sufficiently advance the field to rise to the level of publication in Nature. While a number of interesting substrates were identified, the scale of the study is not sufficiently large to capture the full functionality of the Fe/2OG family or to yield information about their native function. It is also unclear whether it is possible to predict substrate from protein sequence with this library size, which would be a more ambitious and meaningful goal. Instead, the algorithm seems to utilize the dataset to choose library members that are most likely to accept substrates that fall within a similar structural range of the screened library substrates based on a number of chemical structure parameters. I am not able to critically judge this portion, but it seems to me to be a useful tool for analysis of screening data but that the model does yet extend to predict new functions outside the screening range. Therefore in my opinion, I think this is an excellent paper but more suited to a more specialized audience like Nature Chemistry or Nature Chemical Biology as it mainly uses existing methodology to provide a pipeline for screening analysis rather than providing tools for new biochemical or biosynthetic

insight with regard to sequence/function relationships.

(Remarks on code availability)

Referee #3

(Remarks to the Author)
This manuscript introduces a novel strategy for predicting new enzymes suitable for biocatalytic reactions. The proposed methodology initiates with a bioinformatics analysis of the sequence space associated with alpha-ketoglutarate-dependent non-heme iron enzymes, integrated with a chemical space analysis of the substrates pertinent to these enzymes. This innovative approach addresses a critical question in biocatalysis: the feasibility of predicting enzymes for specific reactions without the need for extensive empirical laboratory evaluation and testing. The study is distinguished by its novelty, originality, and significance, and the manuscript is articulated with clarity and precision. Nonetheless, several issues require resolution to strengthen the validity of the approach and enhance the characterization of its applicability and generalizability.

1) A primary consideration is how the current approach compares to traditional methodologies, including random sampling and selection of sequences typically utilized in biochemical research. The described workflow implements a nearest-neighbor search for both enzymes and substrates; however, it remains unclear how effectively the predictions would perform with entirely novel substrates, particularly in light of the limited number of active substrate pairs. Moreover, preliminary calculations based on the provided precision and recall metrics indicate that the predictions may not yet achieve the desired levels of accuracy and efficiency.

2) Within the final panel of enzymes assessed, only a subset was found to be soluble. While predicting the expressibility or solubility of proteins is a significant objective in biocatalysis, it would be beneficial to explore whether categorization or estimation of soluble protein production could be derived from sequence space analysis and clustering. For instance, selectively avoiding proteins from clusters containing enzymes from "exotic" organisms could improve solubility predictions.

3) Figure 5 effectively demonstrates the outcomes of the CATNIP application in biocatalytic reaction discovery, presenting three synthetic examples. The manuscript indicates that the observed reactions primarily involved hydroxylations followed by desaturations. Given the potential for multiple carbon atoms to be hydroxylated, it would be advantageous to provide quantitative estimates regarding the products formed during these reactions.

4) Although the selected enzymes show differences in their overall sequences, they must be quite similar in their three-dimensional structures. It would be beneficial to provide information about the similarities and differences in the active sites of these enzymes.

Minor points:

In Figure 2, the total number of protein sequences listed differs from the number stated in the main manuscript (27005 compared to 27865).

Including recent papers on using machine learning in biocatalysis may be worthwhile. This could encompass the work of researchers such as Damborsky, Arnold, Meilan Huang, and Snajdrova.

(Remarks on code availability)
The code and accompanying description found in the GitHub repository is a valuable resource for readers seeking to grasp the methodology and execute the procedures using a selected substrate of interest. Furthermore, additional details regarding the various substrates and enzymes are accessible, providing an in-depth overview of the data utilized in the training process. This information enhances the understanding of the underlying concepts and facilitates applying the work to specific research needs.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
Evaluation of the revised manuscript 2024-10-22505A-Z
First of all, I am very grateful for the substantial efforts made by the authors to address the reviewer's comments in this revised version of their manuscript, in which Paton et al. describe how they were able to connect "chemical and protein sequence space" exemplified successfully for α-ketoglutarate-dependent non-heme iron (NHI) enzymes.
I appreciate very much that the reviewer's comments motivated them to expand the scope by creating the new enzyme-to-substrate model BioCatSet1 resulting in the new data shown in Fig. 4C and Fig. 5B, and also for adding the enzyme-centric predictive tool to the CATNIP website.
Overall, the manuscript has been substantially improved and I believe that their results are very impressive now.
Indeed, finding many diverse sequences (out of which 70% of the enzymes were uncharacterized, with only 13.7% average

sequence identity and getting 78% expressed and analyzed is indeed an impressive achievement). In addition, the newly identified enzyme-substrate combinations are impressive too notably with the first example of an α-KG dependent enzyme catalyzing oxidative alkene cleavage.
I also appreciate that the preparative reactions were done and proceeded similarly to the analytical scale examples.

Regarding my comments (as reviewer#1), they have been adequately addressed except for these few points:
(i) They have not cited the work by the Reymond group: D. Wahler, F. Badalassi, P. Crotti, J.-L. Reymond, Enzyme fingerprints by fluorogenic and chromogenic substrate arrays, Angew. Chem. Int. Ed. 2001, 40, 4457-4460.

(ii) They wrote in the reply-to-reviewers comments:
Response 3: The reviews by Buller and Lovelock have been added.
This is unfortunately not the case, they are not cited in the revised manuscript! So please add these two citations in a further revision: R. Buller, S. Lutz, R. J. Kazlauskas, R. Snajdrova, J. C. Moore, U. T. Bornscheuer, From nature to industry: Harnessing enzymes for biocatalysis, Science, 382, eadh8615 (2023), DOI: 10.1126/science.adh8615; S. L. Lovelock, R. Crawshaw, S. Basler, C. Levy, D. Baker, D. Hilvert, A. P. Green, The road to fully programmable protein catalysis, Nature 606, 49-58 (2022), DOI: 10.1038/s41586-022-04456-z
Furthermore, in the meantime an update to citation #2 (Wu et al.) has just been published, this new publication should also be cited: Bayer, T., Wu, S. Snajdrova, R., Baldenius K., Bornscheuer, U.T., Angew. Chem. Int. Ed. 64, e202505976 (2025), DOI: 10.1002/anie.202505976

Minor (new) comments:
Figure 1: add the citations for the compounds shown in Fig. 1A (currently Ref. 4, 3, 5,6 and 7,8) to the legend for Fig. 1A. Fig. 1B: add a reference for “GSK2330672”, same for the ATA-r11 example a few lines below (top right reaction shown in Fig. 1B). Also explain the abbreviation “NHI” in the figure legend (it is explained in the main manuscript).
Page 3, top: “…constrained to known reactions discovered through primary or secondary metabolism (Fig. 1B, center).” It is hard to imagine from the figure that the icon/protein structure shown refer to “primary or secondary metabolism” reactions. This could be simply addressed by adding a few lines of text to the legend.
p.4, legend to Figure 2, line 13 (Fig. 2C): should read “Trends in substrates” (plural)
p.5, last sentence before “High throughput biocatalytic reaction discovery”: mention that crude cell lysate from E. coli was used and that no protein purification was done (this is fine with me, but should be mentioned).
p.6, Figure 3A, right box: “32% were reactive”: replace by “32% were converted”?
p.6, Figure 3, legend: what is “c.a.” written in italic font? I think they mean “ca.”
p.6, last paragraph: explain abbreviation “DBU” (compound 10)
p.7, legend to Figure 4, last line: why is “ndcg” written in lower case letters here and not like in the line above (“nCDG”)?
p.9, top paragraph: “you can submit”: better write “one can submit” or “the user can submit”
Chemical structures: all nicely drawn, but I suggest to write “CH3” (with subscript “3”) instead of “Me” for methyl groups. I am not sure what the journal prefers.

Supporting Information:
p. S747: “Luria Broth” should read “Lysogeny Broth” (or use the older name “Luria-Bertani”)
p. S749: legend to Figure S33: better write “The y-axis represents the relative extracted ion count…” or at least write “Y-axis” not “y-axis” as stated.

(Remarks on code availability)

Referee #2

(Remarks to the Author)
The revised manuscript by Paton et al is much improved by the addition of the new application where a given substrate structure can be submitted to a web interface for obtaining a ranked list of enzyme candidates. In this workflow, the authors had quite a good success rate of positive hits, which implies in my opinion that they have fairly good coverage of chemical space for these small to medium-size organic molecules with their library. As such, it is a generally useful tool for this family and helps to define the size of the datasets that may be helpful for this particular enzyme family.

A few minor points:
1. It would be helpful to identify the positions being modified on the substrates (fragmentation analysis etc would be fine) to see if they are binding similarly to what would be expected based on the analysis.
2. I don’t know if it is possible, but is there any way to predict if this size of library (~300 sequences) or so would be general for the characterization of other enzyme families or whether it is specific to the Fe/2OG family.

(Remarks on code availability)

Referee #3

(Remarks to the Author)
Thank you for submitting your revised manuscript and for your thoughtful responses to the reviewers’ critiques. The revisions

have improved the contextualization of your work within the broader landscape of enzyme prediction and biocatalysis, and you have addressed many of the specific points raised, including literature coverage, clarification of methodological scope, and the addition of new data and predictive models.
However, after careful consideration of the reviewers' detailed feedback and your responses, I find that while the manuscript is well-written and the revisions are constructive, some of these broader concerns may warrant further consideration, either in your final revisions or in future work.
- Reviewer 1 and Reviewer 2 both note that the conceptual framework and methodologies employed, while rigorous, closely follow established practices in the field, such as sequence similarity network analysis and machine learning-guided prediction, and similar strategies have been documented previously. The manuscript now provides better context and cites relevant prior work, but the advances, though incremental, do not constitute a sufficiently outstanding conceptual leap.
- The scale of the study, while significant, is not yet broad enough to fully capture the diversity and native functionality of the Fe/2OG enzyme family. The predictive models, though validated with some external data, remain primarily limited to the specific dataset and enzyme class studied. The ability to generalize predictions to new protein families or to provide deeper biochemical insight remains to be demonstrated.
- While the addition of an enrichment metric and clarification of data filtering enhance transparency, questions remain about the predictive power for entirely novel substrates and the practical utility of the models outside the tested scope. Some technical issues, such as product quantification and structural insight, are acknowledged but not fully addressed in the current version.

(Remarks on code availability)
The code provided with the manuscript is a valuable resource for the community. It includes a comprehensive README file with clear instructions for installation and running the application, which greatly facilitates reproducibility. The code is well-organized and, based on my assessment, can be installed and executed as described.

***Reviewer 1***

**Summary:** Paton et al. describe in this interesting and well written manuscript how they attempt to connect “chemical and protein sequence space” exemplified for α-ketoglutarate-dependent non-heme iron (NHI) enzymes. Thus >100 substrates were chosen for a given enzyme and then this was experimentally explored resulting in novel biocatalytic reactions/expanded substrate scopes. To identify the most suitable protein sequence space, they have performed what is common practice now since the Enzyme Function Initiative–Enzyme Similarity Tool was established by John Gerlt. One makes a solid in-depth sequence similarity network (SSN) analysis, which results in a set of (new) sequences encoding different enzymes, which are then studied (from synthetic genes, expressed in E. coli) to identify e.g. what substrate scope these (novel/underexplored/never studied) biocatalysts have. This results in a set of hits and a lot of experimental data, which can then be used for machine learning guided predictions, i.e. identification of positions for mutations to further control the outcome of a biocatalytic reaction (i.e. more selectivity, more stability, broader or smaller substrate scope).

**Critique 1:** Concepts like this have already been studied two decades ago (but at that time SSN etc. was not invented yet). The group of Reymond has created fingerprint profiles for esterase/lipase/epoxide hydrolase etc. using a high-throughput screening system (Wahler et al, cited below), other groups explored error-prone PCR-based mutant libraries (that was easier and cheaper 20 years ago then ordering a set of synthetic genes). The company Diversa published also 20 years ago how to explore sequence space for nitrilases (Robertson et al., cited below). More recently, the group of Pelletier made a major effort to put experimental data from numerous literature-reported reactions catalyzed by the monooxygenase P450-BM3 in a public accessible database to facilitate identification of P450 (variants) with desired features for a given substrate (Fansher et al., cited below), another group has done this for transaminases, where most notably the “chemical space” was already included in the machine learning algorithm to not only compare different enzyme sequences, but also to predict the substrate scope using machine learning (Ao et al., cited below). These are just a few of several examples from literature, where [as stated in their abstract of this manuscript] “strategies for predicting which enzyme is most likely to be compatible with a given small molecule“ have already been developed/described.

None of these papers has been cited here.

Thus, the “longstanding challenge” claimed here in the abstract has already been addressed, at least to some extent.

**Response 1**: We thank the reviewer for highlighting the need for better contextualization of this work with respect to developments in the realm of predicting substrate-enzyme compatibility. Two of the major challenges in developing predictive models are the limited amount of data available and also that previous approaches in predictive model building have focused on highly localized regions of chemical space and/or protein sequence space. A number of studies have focused solely on enzyme activity profiling in local regions of chemical or protein sequence space, which is represented by the work of Reymond, Diversa, and Pelletier. The manuscript introduction has been revised to include a discussion of these types of enzyme activity profiling studies. In addition to addressing the limited data challenge by building a dataset that explores beyond local regions of both chemical and sequence space, a critical advance of this work lies in the generation of a model to navigate across these broad landscapes. This objective requires a significantly greater amount of data and demands different strategies for navigating more broadly than established approaches that make predictions in narrow regions of chemical space and/or protein sequence space. Strategies for making predictions on enzyme activity across panels of variants of a single parent enzyme, like what is detailed in Ao et al. do not translate to making predictions across protein families. In the case cited from Ao, a set of variants is considered with substitutions at seven positions in the protein. This is distinct from making predictions across a protein family with enzymes with an average sequence identity below 14%, as we have done in the work reported here. The introduction has been revised to more clearly communicate these points and provide more context on the accomplishments to date in this field.

**Critique 2**: Another aspect, which I have missed to consider is the correct prediction of enzyme function from sequences in databases, where the Zhao group has established an interesting concept, dubbed ‘contrastive learning’ (Yu et al., cited below). Clearly, this correction of misleading or wrong annotations of protein function must be taken into account when using sequence information to figure out “chemical space” targets.

**Response 2**: The incorrect annotation of sequences has been a significant challenge in identifying suitable biocatalysts for a given type of transformation. The work from Zhao establishing CLEAN is an important and significant advance in annotating enzyme function, in a way that predicts the enzyme commission number that corresponds to the class of reaction an enzyme is anticipated to mediate. This utility is separate from the goal of CATNIP: predicting a specific substrate for a given enzyme.

Please forgive me if I am misinterpreting the comment, but if the concern is how incorrect sequence annotations would impact the enzyme library curation and the subsequent dataset in this study, it is important to note that the library is designed based on the protein family and not the predicted function of a given sequence, which has a higher accuracy that predictions in function. Furthermore, the dataset used to train and test the machine learning models is based exclusively on experimental data, not predicted annotation.

**Critique 3:** Finally, I have missed citations / discussion of two recent reviews relevant in this context (Lovelock et al., Buller et al., cited below). Both reviews also deal in part with the issue of connecting protein sequences to functon(substrate identification/prediction) especially related to the reliable prediction of enzyme properties from protein sequences (which is still challenging using the most advanced AlphaFold version 3.0).

**Response 3:** The reviews by Buller and Lovelock have been added.

**Critique 4:** In summary, the authors have done a great job to establish what they have described in this manuscript. For sure, this is important to other scientists. But I do not see any outstanding novel concept to address the problem of how to connect protein space to chemical space as similar or even identical strategies are documented in literature as exemplified above.

Thus, I can't recommend publication of this work in the journal Nature.

**Response 4:** The revised manuscript more clearly articulates the specific advances provided in this work and also includes greater context for our approach compared to previous work that (1) involves profiling of enzyme activity across substrate and/or enzyme libraries and (2) aims to develop predictive models for navigation between chemical space and protein sequence space. Importantly, the scope of the work reported represents a significant advance in both of these categories, which is more clearly demonstrated through the additional data and predictive tools included in the revised manuscript.

**Critique 5:** There is no need for a major revision (for submission to another journal), I only have a few points:

Ref. 23: it is stated that in 2023 only 1264 journal articles were published on biocatalysis. This must be a completely wrong number. I think this depends on the keywords used for searching (if you search for "electrochemistry", I am confident that this keyword finds most/all publications related to this), i.e. "enzyme catalysis", "enzyme", but also "biotransformation" all must be used to search in the broad area of biocatalysis. I quickly checked this in Web of Science and found 1895 journal articles (not reviews) published in 2023 and I found >20,000 papers when using also the term "enzyme". I suggest to delete Ref. 23 / this statement.

**Response 5:** We have removed this citation.

**Critique 6:** “600,000 fold improvement in yield” stated on p.3 for the research described in Ref. 30: Yield is isolated product, how can one improve this 600,000 fold as it is already a challenge to isolate (find) 0,1% of a product. Is this really yield or just rate acceleration? Please check the term “yield” before such a statement is written in the manuscript.

**Response 6:** We thank the Reviewer for raising this point, and we have changed the term “yield” to “cumulative activity.”

**Critique 7:** A side note: The authors also developed the open access interface “Catnip”, which should help to use their predictive workflow. I very much appreciate this effort. I have tried Catnip drawing a simple molecule (e.g. acetophenone) and would then expect that the first two hits would be “ketoreductase” to yield the chiral alcohol or “amine transferase” to yield the chiral amine (and maybe a list of various wildtype enzymes or some mutants or acetophenone derivatives with some substitutions at the aromatic ring). The outcome was confusing to me, as numerous highly complex (or unrelated) molecules were predicted by Catnip, but none of them related at all to the expected (most logical first enzymatic steps) suggestion of enzymes or ‘chemical space’. Also, the list of enzymes in the ranking made no sense to me.

**Response 7:** We appreciate that you navigated to catnip.cheme.cmu.edu to test the user interface for enzyme predictions. However, as stated in the manuscript, CATNIP is an open access web platform which is founded on BioCatSet1, a dataset that was generated wholly by the exploration of α-ketoglutarate non-heme iron-dependent enzymes. Therefore, CATNIP is currently restricted to this superfamily of proteins, and should not generate predictions that extend to other protein families such as ketoreductases or amine transaminases. In light of this confusion, the authors have provided additional clarification within the manuscript that the substrate-to-enzyme CATNIP model is specific to α-ketoglutarate non-heme iron-dependent enzymes.

**Summary:** We thank Reviewer 1 for their feedback that has helped us improve the presentation of work in this manuscript.

***Reviewer 2***

**Summary:** The manuscript by Paton et al describes an interesting study where the authors utilize EFI tools to select a broad range of 314 Fe/2OG-dependent enzymes for screening against a panel of 111 substrates. The substrates ranged from small aromatics and amino acids to larger extended and polycyclic structures to around ~500 Da in size. From this screen, a good number of hits were found with 32% of substrates being accepted by an enzyme and 38% of enzymes showing competence on a substrate, yielding 215 new reactions. Machine learning was then applied using this dataset as well as literature reactions (139 reactions) with the goal of predicting enzymes to screen for specific substrates. This approach was applied to three different test molecules, for which transformations were successfully obtained.

**Critique 1:** Overall, I enjoyed reading the manuscript and I think that the study is well-designed and well-conceived. However, I am not sure that the methods or results sufficiently advance the field to rise to the level of publication in Nature. While a number of interesting substrates were identified, the scale of the study is not sufficiently large to capture the full functionality of the Fe/2OG family or to yield information about their native function.

**Response 1:** We thank the Reviewer for this positive assessment of our work, and we appreciate the questions of whether this predictive tool could be applied to substrates beyond the dataset or to provide insight on the function of Fe/2OG enzymes in this vast protein family. These questions motivated us to dramatically expand the scope of this project to encompass the originally reported substrate-to-enzyme model and a new enzyme-to-substrate model. In regards to the scale of the BioCatSet1 and aKGLib1 datasets, aKGLib1 offers the most broadly encompassing report of reactivity profiling across wild-type α-ketoglutarate non heme iron-dependent enzymes to date. Furthermore, the high-throughput reaction profiling campaign makes up ~35,000 novel substrate-enzyme pairs. Although only a sliver of these profiled enzyme/substrate pairs displayed compatibility, the scoring of the models is dependent on the positive (hits) and negative data. This allows the model to make predictions on substrates outside of the BioCatSet1, providing a general tool for synthetic chemists looking for an enzyme that might work with a given target substrate. Further, the newly built enzyme-to-substrate model (see Fig. 4C and Fig. 5B) allows one to gain insight on the area of chemical space an enzyme outside aKGLib1 might operate within. This enzyme-centric predictive tool has been added to the CATNIP website (catnip.cheme.cmu.edu).

**Critique 2:** It is also unclear whether it is possible to predict substrate from protein sequence with this library size, which would be a more ambitious and meaningful goal. Instead, the algorithm seems to utilize the dataset to choose library members that are most likely to accept substrates that fall within a similar structural range of the screened library substrates based on a number of chemical structure parameters.

**Response 2:** Based on this insightful feedback, we were motivated to tackle this ambitious goal. Therefore, we used the BioCatSet1 dataset to build and train an enzyme-to-substrate model, capable of predicting compatible substrates for given enzyme sequences (Fig. 5B). This model is focused on α-ketoglutarate non-heme iron-dependent enzymes, but the strategy is anticipated to translate to additional protein families, and we are actively pursuing this direction.

**Critique 3:** I am not able to critically judge this portion, but it seems to me to be a useful tool for analysis of screening data but that the model does yet extend to predict new functions outside the screening range.

**Response 3:** This critique raises a good question, but we are pleased to share that the models have been productive at generating predictions outside of the existing dataset. We have added the experimental validation of substrates and enzymes outside the dataset in Figure 5. Notably, there are examples of meaningful predictions with substrates and enzymes in the training set, test set, and through external validation.

**Summary:** Therefore in my opinion, I think this is an excellent paper but more suited to a more specialized audience like Nature Chemistry or Nature Chemical Biology as it mainly uses existing methodology to provide a pipeline for screening analysis rather than providing tools for new biochemical or biosynthetic insight with regard to sequence/function relationships.

**Summary Response:** We thank Reviewer 2 for their questions that have motivated us to build and validate an enzyme-to-substrate model and include additional data to highlight the application of these predictive models beyond substrates and enzymes contained within the dataset. We anticipate that the revised manuscript will be perceived as more impactful based on these additions.

***Reviewer 3***

**Summary:** This manuscript introduces a novel strategy for predicting new enzymes suitable for biocatalytic reactions. The proposed methodology initiates with a bioinformatics analysis of the sequence space associated with alpha-ketoglutarate-dependent non-heme iron enzymes, integrated with a chemical space analysis of the substrates pertinent to these enzymes. This innovative approach addresses a critical question in biocatalysis: the feasibility of predicting enzymes for specific reactions without the need for extensive empirical laboratory evaluation and testing. The study is distinguished by its novelty, originality, and significance, and the manuscript is articulated with clarity and precision. Nonetheless, several issues require resolution to strengthen the validity of the approach and enhance the characterization of its applicability and generalizability.

**Critique 1:** A primary consideration is how the current approach compares to traditional methodologies, including random sampling and selection of sequences typically utilized in biochemical research. The described workflow implements a nearest-neighbor search for both enzymes and substrates; however, it remains unclear how effectively the predictions would perform with entirely novel substrates, particularly in light of the limited number of active substrate pairs. Moreover, preliminary calculations based on the provided precision and recall metrics indicate that the predictions may not yet achieve the desired levels of accuracy and efficiency.

**Response 1:** In order to compare the substrate-centric model to previous methods, we have added the use of an enrichment metric. This metric compares the ability of the model to effectively predict compatible substrates and enzymes as compared to randomly sampling aKGLib1. This metric has been included for both the substrate-to-enzyme and enzyme-to-substrate baseline and gradient boosted models in Fig. 4C. We believe that this is the most unbiased metric available to compare the performance of the models and random sampling. In addition, the dataset status of substrates and enzymes has been indicated in the revised manuscript to highlight examples of substrates and enzymes external to those contained within BioCatSet1 and aKGLib1. We have found that substrates and enzymes outside of the dataset are compatible with the developed predictive models.

**Critique 2:** Within the final panel of enzymes assessed, only a subset was found to be soluble. While predicting the expressibility or solubility of proteins is a significant objective in biocatalysis, it would be beneficial to explore whether categorization or estimation of soluble protein production could be derived from sequence space analysis and clustering. For instance, selectively avoiding proteins from clusters containing enzymes from "exotic" organisms could improve solubility predictions.

**Response 2:** We agree that filtering enzymes by predicting solubility or expression would have been a wise approach for the library curation. If we were to redesign this project, we would have used a computational tool such as EnzymeMiner (Hon, et al., Nucleic Acids Research, 2020, Vol. 48, Web Server issue) to potentially increase the percentage of library enzymes that can be obtained in soluble form in the high throughput workflow used. We plan to include this step in future work focused on building enzyme libraries in different protein families. Without this type of filter in place, 70% of the sequences within AKGLib1 produced soluble enzyme as observed by SDS-PAGE and numerous enzymes for which a soluble enzyme band was not clearly observed, still catalyzed biocatalytic reactions, suggesting that the percentage of soluble enzymes was higher than the approximation based on SDS-PAGE as a readout.

**Critique 3:** Figure 5 effectively demonstrates the outcomes of the CATNIP application in biocatalytic reaction discovery, presenting three synthetic examples. The manuscript indicates that the observed reactions primarily involved hydroxylations followed by desaturations. Given the potential for multiple

carbon atoms to be hydroxylated, it would be advantageous to provide quantitative estimates regarding the products formed during these reactions.

**Response 3:** We agree with the Reviewer that there are additional layers to consider beyond reactivity between a substrate/enzyme pair, with selectivity and reaction type as important examples. To develop a predictive model that can tackle these questions, we will need to collect data that will allow for these types of predictions. In the context for C–H functionalization reactions, it is difficult to characterize and quantify all of the possible products in a high throughput fashion. Thus, to develop models that can address these additional layers, we are focusing on reactions that have a more limited set of potential products, and we look forward to reporting these results in due course.

**Critique 4:** Although the selected enzymes show differences in their overall sequences, they must be quite similar in their three-dimensional structures. It would be beneficial to provide information about the similarities and differences in the active sites of these enzymes.

**Response 4:** We agree that question of structural diversity present within aKGLib1 is very interesting given the sequences on average only share 14% sequence identity. To gain some insight on the structural differences, we used FoldSeek (van Kempen, et al. Nat Biotechnol, 2024, 42, 243–246) to perform a multiple structure alignment on a subset of the enzyme library (n = 190). The local distance difference test resulted in a score of 0.189 and minimal structural overlap was observed visually. Of note, most of the structures used in this modelling effort were generated by AlphaFold2 and this score reflects the structural diversity in the entire structure. Based on the complexities of making structural comparisons, we have not included this analysis with this work as our model only considers the sequence. However, we are very interested in developing models based on enzyme structure and are currently pursuing these directions separately.

**Critique 5:** In Figure 2, the total number of protein sequences listed differs from the number stated in the main manuscript (27005 compared to 27865).

**Response 5:** This inconsistency is a result of the enzyme filtering process detailed with the ‘Sequence similarity network generation and enzyme selection’ section of the Supplementary Information. In order to avoid confusion, we have included the pre-filtered and post-filtered values within the main text of the revised manuscript.

**Critique 6:** Including recent papers on using machine learning in biocatalysis may be worthwhile. This could encompass the work of researchers such as Damborsky, Arnold, Meilan Huang, and Snajdrova.

**Response 6:** We agree with the Reviewer regarding the need to discuss and reference additional biocatalytic machine learning models within this publication. In light of this, we have included a new paragraph within the introduction to provide context in this field and highlight the work of Ao, Yu, Wahler, Robertson, Fansher, Baker, Arnold, Snajdrova, and Lercher.

**Critique 7:** The code and accompanying description found in the GitHub repository is a valuable resource for readers seeking to grasp the methodology and execute the procedures using a selected substrate of interest. Furthermore, additional details regarding the various substrates and enzymes are accessible, providing an in-depth overview of the data utilized in the training process. This information enhances the understanding of the underlying concepts and facilitates applying the work to specific research needs.

**Response 7:** We thank the Reviewer for appreciating the work we have put into detailing our findings and making these models accessible to the community.

## Referees' comments:

### Referee 1

**Comment 1.1:** First of all, I am very grateful for the substantial efforts made by the authors to address the reviewer's comments in this revised version of their manuscript, in which Paton et al. describe how they were able to connect "chemical and protein sequence space" exemplified successfully for α-ketoglutarate-dependent non-heme iron (NHI) enzymes.

I appreciate very much that the reviewer's comments motivated them to expand the scope by creating the new enzyme-to-substrate model BioCatSet1 resulting in the new data shown in Fig. 4C and Fig. 5B, and also for adding the enzyme-centric predictive tool to the CATNIP website. Overall, the manuscript has been substantially improved and I believe that their results are very impressive now.

Indeed, finding many diverse sequences (out of which 70% of the enzymes were uncharacterized, with only 13.7% average sequence identity and getting 78% expressed and analyzed is indeed an impressive achievement). In addition, the newly identified enzyme-substrate combinations are impressive too notably with the first example of an α-KG dependent enzyme catalyzing oxidative alkene cleavage.

I also appreciate that the preparative reactions were done and proceeded similarly to the analytical scale examples.

**Response 1.1:** We thank Referee 1 for their thoughtful and positive assessment of the revised manuscript.

**Comment 1.2:** Regarding my comments, they have been adequately addressed except for these few points:

They have not cited the work by the Reymond group: D. Wahler, F. Badalassi, P. Crotti, J.-L. Reymond, Enzyme fingerprints by fluorogenic and chromogenic substrate arrays, Angew. Chem. Int. Ed. 2001, 40, 4457-4460.

**Response 1.2:** The Wahler paper has been included as reference 19 in the revised manuscript.

**Comment 1.3:** They wrote in the reply-to-reviewers comments: "The reviews by Buller and Lovelock have been added." This is unfortunately not the case, they are not cited in the revised manuscript! So please add these two citations in a further revision: R. Buller, S. Lutz, R. J. Kazlauskas, R. Snajdrova, J. C. Moore, U. T. Bornscheuer, From nature to industry: Harnessing enzymes for biocatalysis, Science, 382, eadh8615 (2023), DOI: 10.1126/science.adh8615; S. L. Lovelock, R. Crawshaw, S. Basler, C. Levy, D. Baker, D. Hilvert, A. P. Green, The road to fully programmable protein catalysis, Nature 606, 49-58 (2022), DOI: 10.1038/s41586-022-04456-z.

**Response 1.3:** We thank the Reviewer for bringing this to our attention, as these references were accidentally left out of the previous revision. These citations are now included as references 7 and 14, respectively.

**Comment 1.4:** Furthermore, in the meantime an update to citation #2 (Wu et al.) has just been published, this new publication should also be cited: Bayer, T., Wu, S. Snajdrova, R., Baldenius K., Bornscheuer, U.T., Angew. Chem. Int. Ed. 64, e202505976 (2025), DOI: 10.1002/anie.202505976

**Response 1.4:** This updated citation is reference 3 in the revised manuscript.

**Comment 1.5:** Minor (new) comments: Figure 1: add the citations for the compounds shown in Fig. 1A (currently Ref. 4, 3, 5,6 and 7,8) to the legend for Fig. 1A. Fig. 1B: add a reference for "GSK2330672", same for the ATA-r11 example a few lines below (top right reaction shown in Fig. 1B). Also explain the abbreviation "NHI" in the figure legend (it is explained in the main manuscript).

**Response 1.5:** Due to space constraints, the original Figure 1A has been removed from the manuscript. The other suggested changes have been made.

**Comment 1.6:** Page 3, top: "…constrained to known reactions discovered through primary or secondary metabolism (Fig. 1B, center)." It is hard to imagine from the figure that the icon/protein structure shown refer to "primary or secondary metabolism" reactions. This could be simply addressed by adding a few lines of text to the legend.

**Response 1.6:** The reference to Figure 1 has been removed from this sentence.

**Comment 1.7:** p.4, legend to Figure 2, line 13 (Fig. 2C): should read "Trends in substrates" (plural)

**Response 1.7:** This typo has been corrected.

**Comment 1.8:** p.5, last sentence before "High throughput biocatalytic reaction discovery": mention that crude cell lysate from E. coli was used and that no protein purification was done (this is fine with me, but should be mentioned).

**Response 1.8:** This change has been made.

**Comment 1.9:** p.6, Figure 3A, right box: "32% were reactive": replace by "32% were converted"?

**Response 1.9:** This change has been made.

**Comment 1.10:** p.6, Figure 3, legend: what is "c.a." written in italic font? I think they mean "ca."

**Response 1.10:** This change has been made.

**Comment 1.11:** p.6, last paragraph: explain abbreviation "DBU" (compound 10)

**Response 1.11:** This change has been made.

**Comment 1.12:** p.7, legend to Figure 4, last line: why is "ndcg" written in lower case letters here and not like in the line above ("nCDG")?

**Response 1.12:** This typo has been corrected.

**Comment 1.13:** p.9, top paragraph: "you can submit": better write "one can submit" or "the user can submit"

**Response 1.13:** This change has been made.

**Comment 1.14:** Chemical structures: all nicely drawn, but I suggest to write “CH3” (with subscript “3”) instead of “Me” for methyl groups. I am not sure what the journal prefers.

**Response 1.14:** As a stylistic preference, we have retained the use of “Me” labels for methyl groups.

**Comment 1.15:** Supporting Information: p. S747: “Luria Broth” should read “Lysogeny Broth” (or use the older name “Luria-Bertani”)

**Response 1.15:** This change has been made.

**Comment 1.16:** p. S749: legend to Figure S33: better write “The y-axis represents the relative extracted ion count…” or at least write “Y-axis” not “y-axis” as stated.

**Response 1.16:** This change has been made.

**Referee #2**

**Comment 2.1:** The revised manuscript by Paton et al is much improved by the addition of the new application where a given substrate structure can be submitted to a web interface for obtaining a ranked list of enzyme candidates. In this workflow, the authors had quite a good success rate of positive hits, which implies in my opinion that they have fairly good coverage of chemical space for these small to medium-size organic molecules with their library. As such, it is a generally useful tool for this family and helps to define the size of the datasets that may be helpful for this particular enzyme family.

**Response 2.1:** We thank Referee 2 for their positive assessment of the revised manuscript.

**Comment 2.2:** A few minor points: It would be helpful to identify the positions being modified on the substrates (fragmentation analysis etc would be fine) to see if they are binding similarly to what would be expected based on the analysis.

**Response 2.2:** We are very interested in the question of site-selectivity. This report of a predictive model is focused on the question of substrate-enzyme compatibility. Future efforts will be focused on additional layers of selectivity, including the site of the reaction.

**Comment 2.3:** I don’t know if it is possible, but is there any way to predict if this size of library (~300 sequences) or so would be general for the characterization of other enzyme families or whether it is specific to the Fe/2OG family.

**Response 2.3:** This is an interesting question. Based on our work within this project and others, I anticipate that there is not a fixed library size that is ideal, but rather, that this will depend on the individual enzyme class and the reaction of interest.

**Referee #3**

**Comment 3.1:** Thank you for submitting your revised manuscript and for your thoughtful responses to the reviewers’ critiques. The revisions have improved the contextualization of your work within the broader landscape of enzyme prediction and biocatalysis, and you have addressed many of the specific points raised, including literature coverage, clarification of methodological scope, and the addition of new data and predictive models.

However, after careful consideration of the reviewers' detailed feedback and your responses, I find that while the manuscript is well-written and the revisions are constructive, some of these broader concerns may warrant further consideration, either in your final revisions or in future work.

**Response 3.1:** We appreciate the overall positive feedback from Referee 3 on the revised manuscript and will keep in mind the points raised in considering future work.

**Comment 3.2:** Reviewer 1 and Reviewer 2 both note that the conceptual framework and methodologies employed, while rigorous, closely follow established practices in the field, such as sequence similarity network analysis and machine learning-guided prediction, and similar strategies have been documented previously. The manuscript now provides better context and cites relevant prior work, but the advances, though incremental, do not constitute a sufficiently outstanding conceptual leap.

**Response 3.2:** In the assessment of the revised manuscript, both Reviewers 1 and 2 were satisfied with the changes made to the framing of the work and the advances provided by this work. For example, Reviewer 1 commented, "Overall, the manuscript has been substantially improved and I believe that their results are very impressive now." And Reviewer 2 commented on the scope saying, "The revised manuscript by Paton et al is much improved by the addition of the new application where a given substrate structure can be submitted to a web interface for obtaining a ranked list of enzyme candidates." Thus, the concerns originally raised by the other Reviewers have been addressed to their satisfaction.

**Comment 3.3:** The scale of the study, while significant, is not yet broad enough to fully capture the diversity and native functionality of the Fe/2OG enzyme family. The predictive models, though validated with some external data, remain primarily limited to the specific dataset and enzyme class studied. The ability to generalize predictions to new protein families or to provide deeper biochemical insight remains to be demonstrated.

**Response 3.3:** It is true that the predictive model generated is specific to the Fe/2OG enzyme family, however, the other statements made require some clarification. The predictive models are not limited to substrates and Fe/2OG enzymes within the dataset as detailed in the manuscript. We anticipate that this two-pronged approach- high throughput experimentation coupled with model building- can be successfully applied to additional protein families.

**Comment 3.4:** While the addition of an enrichment metric and clarification of data filtering enhance transparency, questions remain about the predictive power for entirely novel substrates and the practical utility of the models outside the tested scope. Some technical issues, such as product quantification and structural insight, are acknowledged but not fully addressed in the current version.

**Response 3.4:** We acknowledge the Reviewer's comments, however, sufficient detail is not included that would allow us to address their concerns.

**Comment 3.5:** The code provided with the manuscript is a valuable resource for the community. It includes a comprehensive README file with clear instructions for installation and running the application, which greatly facilitates reproducibility. The code is well-organized and, based on my assessment, can be installed and executed as described.

**Response 3.5:** We thank the Referee for acknowledging the quality of the code associated with this manuscript.