Article

Connecting chemical and protein sequence space to predict biocatalytic reactions

https://doi.org/10.1038/s41586-025-09519-5

Received: 21 October 2024

Accepted: 12 August 2025

Published online: 1 October 2025

Open access

Check for updates

Alexandra E. Paton[1], Daniil A. Boiko[2], Jonathan C. Perkins[1,3], Nicholas I. Cemalovic[1,3], Thiago Reschützegger[4], Gabe Gomes[2,5,6,7]✉ & Alison R. H. Narayan[1,3]✉

The application of biocatalysis in synthesis has the potential to offer streamlined routes towards target molecules[1], tunable catalyst-controlled selectivity[2], as well as processes with improved sustainability[3]. Despite these advantages, biocatalysis is often a high-risk strategy to implement, as identifying an enzyme capable of performing chemistry on a specific intermediate required for a synthesis can be a roadblock that requires extensive screening of enzymes and protein engineering to overcome[4]. Strategies for predicting which enzyme and small molecule are compatible have been hindered by the lack of well-studied biocatalytic reaction datasets[5]. The underexploration of connections between chemical and protein sequence space constrains navigation between these two landscapes. Here we report a two-phase effort relying on high-throughput experimentation to populate connections between productive substrate and enzyme pairs and the subsequent development of a tool, CATNIP, for predicting compatible α-ketoglutarate (α-KG)/Fe(II)-dependent enzymes for a given substrate or, conversely, for ranking potential substrates for a given α-KG/Fe(II)-dependent enzyme sequence. We anticipate that our approach can be readily expanded to further enzyme and transformation classes and will derisk the investigation and application of biocatalytic methods.

The use of enzymes in small-molecule synthesis has transformed the production of commodity chemicals and enabled the construction of complex molecules for decades[6]. Recent examples of biocatalytic routes to achieve the commercial production of pharmaceutical agents underscore the potential of designing synthetic strategies that include key biocatalytic steps[3]. New enzyme-mediated process routes towards drugs have decreased step counts by 33% and more than doubled overall yield, on average, compared with the highest-performing chemical syntheses[7] (Supplementary Information, scheme 1). The potential to enable new routes towards target molecules through biocatalysis exists in arenas outside process chemistry, as evidenced by a growing body of work from academic groups specializing in chemoenzymatic synthesis[1] and the potential of enzymatic late-stage functionalization in discovery chemistry[2].

One common limitation of biocatalysis is the unpredictable substrate scope of individual enzymes, which can contribute to difficulty in developing a biocatalytic method[4]. Even simple methyl, ethyl and propyl substituent series, which typically do not show notable reactivity differences for small-molecule catalysts, can reveal large differences in an enzymatic reaction[8]. Therefore, to plan an enzymatic reaction into a synthetic route carries substantial risk if the exact reaction on the planned substrate is not already known. As a result, the application of biocatalysis is often constrained to known reactions discovered through primary or secondary metabolism[9]. Once a starting point enzyme–substrate pair is identified, local exploration of chemical space or protein sequence space can lead to the desired reactivity. For example, towards the synthesis of GSK2330672, the known transformation of epichlorohydrin by an epoxide hydrolase was used as a starting point for local chemical space exploration to extend this chemistry to a new epoxide[10] (Fig. 1a). As an alternative, we can explore local protein sequence space through protein engineering[11]. Several notable examples of protein engineering have been applied in the synthesis of pharmaceutical agents[7,12], including the engineering of a transaminase for the synthesis of sacubitril, a neprilysin inhibitor, which involved the substitution of 26 amino-acid residues to achieve a 500,000-fold improvement in activity[13] (Fig. 1a). Thus, established strategies for applying biocatalysis rely heavily on known reactions and local exploration in chemical space and protein sequence space forward from these defined connections (Fig. 1b). Unfortunately, the percentage of enzymes for which chemistry has been experimentally characterized is extremely low, with less than 0.3% of sequenced enzymes having a computationally annotated function[5]. As such, most enzymes do not have known connections to substrate chemical space, contributing to the difficulty in tapping into the potential that these catalysts could bring to the scientific community.

Machine learning methods can expedite the biocatalytic reaction discovery process[14]. For example, a contrastive learning model was developed to predict the enzyme commission number of uncharacterized enzymes[15]. This provides a prediction for what type of reaction a given enzyme is capable of. However, it does not guide us towards the native substrate of an enzyme nor provide information on the substrate

[1]Life Sciences Institute, University of Michigan, Ann Arbor, MI, USA. [2]Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA. [3]Department of Chemistry, University of Michigan, Ann Arbor, MI, USA. [4]Department of Chemical Engineering, Federal University of Santa Maria, Santa Maria, Brazil. [5]Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA, USA. [6]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA. [7]Wilton E. Scott Institute for Energy Innovation, Carnegie Mellon University, Pittsburgh, PA, USA. ✉e-mail: gabegomes@cmu.edu; arhardin@umich.edu

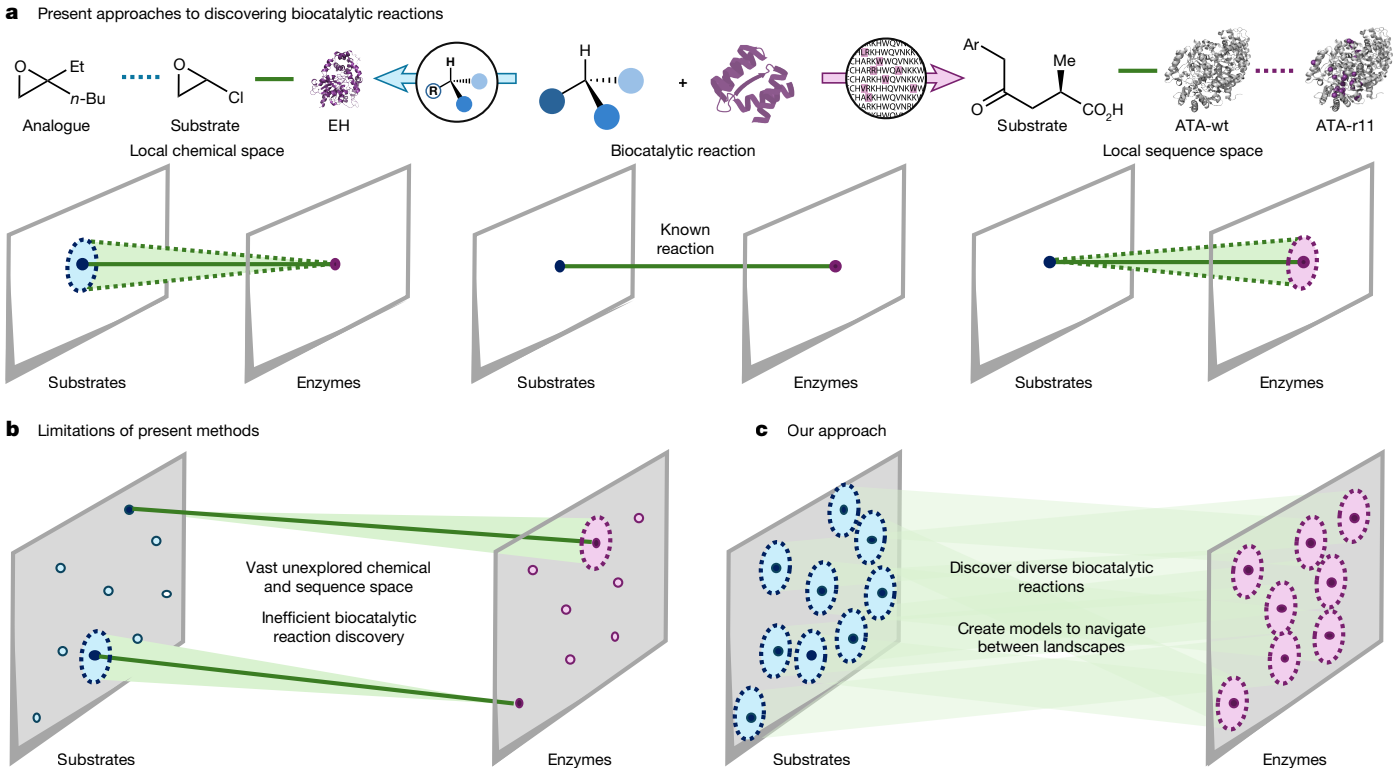

Fig. 1 | Present state of the art in biocatalytic reaction discovery. **a**, Established methods for new biocatalytic reaction discovery. Known connections between chemical and protein sequence space can be exploited for new reaction discovery through local exploration. The known reaction between epichlorohydrin and epoxide hydrolase (EH) was used to enable the reaction on the epoxide analogue towards the synthesis of GSK2330672 (ref. 10). Alternatively, local protein sequence space was explored through protein engineering to improve the transformation of the known substrate (Ar = *p*-biphenyl) with wild-type (wt) amino transaminase (ATA), resulting in ATA-r11 after 11 rounds of directed evolution (positions of mutations shown in purple)[13]. **b**, Limitations of present methods. Expansion of characterized biocatalytic reactivity is limited to local exploration of chemical and sequence space, inhibiting larger, non-intuitive leaps between the landscapes. There remains a vast unexplored region of substrates and enzymes with unknown biocatalytic reactivity, creating a higher risk for their incorporation as key steps in chemical synthesis. There is at present no method to predict compatible enzymes or substrates in the NHI enzyme superfamily. **c**, Our approach to streamline biocatalytic reaction discovery. We examined diverse substrates and protein sequences for new biocatalytic reactions and use these data to build machine learning models to predict compatible enzymes and substrates.

scope. Computational tools can also predict other qualities of a protein based on its sequence, such as EnzymeMiner, which predicts amenability to heterologous expression in *Escherichia coli* (*E. coli*)[16]. These tools are useful for guiding initial selection of enzymes to explore experimentally. However, it is well documented that enzyme annotation and predicted compatible substrates often do not align with experimental validation[17], hindering their use in generating in silico datasets.

Advances towards achieving predictability in biocatalysis have provided solutions for navigating locally in either chemical or sequence space. Reaction discovery campaigns, such as the profiling of a nitrilase library against a small panel of highly similar substrates[18] and the fluorogenic-guided investigation of the substrate scope of hydrolases[19], have revealed trends in enzyme promiscuity. There has also been work to curate datasets detailing the substrate scope of variants of a given parent enzyme[20], as exemplified by the profiling of P450 BM3 variants against a panel of small molecules[21]. From these types of dataset, local sequence space exploration can be aided by machine learning tools to identify variant enzymes with superior catalytic activity[22], stereoselectivity[23], substrate scope[24] and thermal stability[25]. These datasets explore locally in sequence space, chemical space or both landscapes (Fig. 1b). Attempts to predict biocatalytic reactions have been carried out in several enzyme families[26,27]. However, established approaches have limitations in applicability and accessibility. These constraints include difficulty extrapolating beyond the training set[28] and an absence of user-friendly tools[26]. Further, established approaches that are not restricted to a specific enzyme family underperform[28], probably because of the large differences in substrate selectivity observed across various protein families[29]. The generation of models that rely on pre-existing datasets[30], which are largely taken from the biosynthetic and metabolism literature, create limitations, as the array of enzymes and substrates have not been experimentally validated against each other, leading to a risk of false negatives, poor annotations and inaccurate proposed biocatalytic reactions that can lead to false positives[17].

Towards characterizing the chemistry possible across an enzyme family and derisking the incorporation of biocatalytic steps into synthetic routes, we visualized a two-pronged approach involving high-throughput experimentation and machine learning. We anticipated that this would require conducting reactions that profile substrates sampled across chemical space with enzymes that represent the sequence diversity encompassed by a protein family (Fig. 1c). Once a sufficient dataset was obtained, then machine learning models could be built to navigate between these two landscapes and enable the discovery of biocatalytic reactions in a substrate-oriented or enzyme-oriented fashion. Here we detail the first, to our knowledge, example of this approach focused on C–H functionalization reactions mediated by α-KG non-haem iron(II)-dependent (NHI) enzymes. Ultimately, this experimental effort led to the discovery of more than 200 biocatalytic reactions and provided the data necessary to build a web-based toolkit to suggest compatible substrates and enzymes for an oxidative biocatalytic transformation.

In considering which protein family and reaction class to use as a test case, we sought an enzyme class that has proved to be useful on the preparative scale[31] and performs valuable reactions while arguably still underdeveloped in synthetic chemistry[32]. On the basis of this, we chose to focus on a subclass of NHI enzymes that use α-KG as a co-substrate. From a reactivity standpoint, this class of enzymes is appealing on the basis of their ability to access a range of chemistries from a conserved radical intermediate to afford C–H functionalization products and desaturation products and mediate skeletal rearrangements[33,34] (Fig. 2a). Also, α-KG-dependent NHI enzymes have practical advantages over other types of enzymes that can mediate the cleavage of strong C–H bonds or perform oxidative transformations on several bonds. For example, other subclasses of NHI and cytochrome P450 enzymes are fuelled by electrons that are often supplied by a partner reductase[35], whereas α-KG-dependent NHI enzymes rely on the oxidation of the small-molecule co-substrate α-KG to drive the formation of the active oxidant species[36] (Fig. 2a). This difference provides a more uniform set of conditions for α-KG-dependent NHI reactions, which have proved to be scalable[37].

To design a library of α-KG-dependent NHI enzymes that represent the sequence diversity of this protein family, we gathered all sequences annotated to have the facial triad of iron-coordinating residues that is conserved for hydroxylases[38] (Fig. 2b). Using the Enzyme Function Initiative–Enzyme Similarity Tool (EFI-EST), 265,632 unique sequences were associated with this class[39]. To reduce the number of sequences to a manageable amount, redundant orthologues (>90% similarity) and clusters containing enzymes associated with primary metabolism were removed, giving a resultant sequence similarity network (SSN) consisting of 27,005 sequences (Fig. 2b). Work by Lewis and colleagues demonstrated that SSN representations can reveal trends in sequence–substrate relationships within the flavin-dependent halogenase family[40] and subsequent studies have shown this correlation in further enzyme classes[41–45]. Therefore, we sampled several clusters as a strategy to achieve a protein library with a broad substrate scope. In total, 102 sequences were selected from the most populated cluster, 125 uncharacterized sequences from poorly annotated clusters and 87 further sequences of enzymes with known or proposed function were selected to arrive at a 314 enzyme library (aKGLib1; Supplementary Fig. 1).

Of the enzymes selected, 94 (30%) have a known or suggested native reaction including hydroxylation, desaturation, halogenation, epoxidation, endoperoxidation, demethylation, C–C bond formation and skeletal rearrangements (Fig. 2c and Supplementary Figs. 2 and 3). In an attempt to extrapolate beyond known activity, we used the enzyme commission machine learning model CLEAN (Contrastive Learning enabled Enzyme ANnotation)[15]. CLEAN assigned enzymes in aKGLib1 as oxidoreductases, transferases, hydrolases, lyases and isomerases, of which 80% were annotated with low confidence (Supplementary Table 1 and Supplementary Figs. 4 and 5). As anticipated, trends in substrate class are evident from the SSN generated at a more stringent alignment score threshold (Fig. 2c). For example, at alignment score 75, enzymes characterized to be compatible with an indolizidine scaffold are found within a cluster. All selected sequences showed an average sequence percent identity of 13.7%, indicating high library sequence diversity (Fig. 2c). DNA for the library was synthesized and cloned into a pET-28b(+) expression vector. *E. coli* cells were transformed with plasmids encoding for each library member and overexpression was carried out in 96-well-plate format. Sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) analysis of crude cell lysate showed clear protein bands at the expected molecular weight for 78% of enzymes (Supplementary Figs. 6–29).

High-throughput biocatalytic reaction discovery

With aKGLib1 in hand, we next investigated the reactivity of each enzyme in a high-throughput fashion. To profile the reactivity of each enzyme with a range of substrates, reactions were performed on the 50-µl scale in 96-well plates and were analyzed by liquid chromatography-mass spectrometry (LC-MS). Each reaction was investigated for masses corresponding to hydroxylation, chlorination, desaturation and rearrangement products (Fig. 3a). Notably, each reaction was conducted in triplicate and compared with two negative controls: (1) cell pellet containing no overproduced NHI enzyme and (2) a no-substrate control.

More than a hundred compounds were assessed as substrates for each enzyme in aKGLib1, including a range of scaffolds from commercially available amino acids to drugs and other complex molecules (Supplementary Table 5). Of the 111 substrates evaluated in reactions with the entire enzyme library, 35 compounds were transformed by at least one enzyme in aKGLib1, a 32% success rate (Fig. 3a). Furthermore, 119 of 314 enzymes showed biocatalytic activity on at least one substrate tested, including 74 with no previously reported activity (Supplementary Table 4). Notably, numerous enzymes for which a protein band was not clearly seen by SDS-PAGE analysis showed biocatalytic activity. Most of the observed reactions were hydroxylation, although desaturation reactions constituted about 20% of the reactions discovered (Fig. 3a). In total, 215 new biocatalytic reactions were observed (Supplementary Table 3). The collection of discovered reactions encompassed substrates that varied substantially in structure, including natural products such as cannabidivarin (**4**), humulene (**12**) and harmaline (**13**), chemical building blocks such as cinnamic acid analogue (**1**) and usnic acid (**3**), common reagents such as 1,8-diazabicyclo[5.4.0]undec-7-ene (DBU, **10**) and pharmaceutical agents (for example, glyburide (**5**)). To define the relationship between compounds, each substrate was quantified with MORFEUS descriptors[46] to generate 21 parameters including measurements of sterics (for example, volume, solvent-accessible surface area), electronics (for example, HOMO and LUMO energies, electrophilicity) and intermolecular interactions (for example, dispersion descriptors, charge) (Supplementary Table 7). With these features quantified, we carried out a principal component analysis to represent the compounds in chemical space (Fig. 3b). Notably, substrates transformed by enzymes in aKGLib1 are well dispersed in chemical space. Ultimately, our experimental effort produced hundreds of new connections between chemical space and protein sequence space. Thus, the development of models that would allow for navigation between these landscapes was now within reach.

Translation from reaction data to machine learning models

With the goal of creating robust compatibility predictive models, we sought to maximize the number of biocatalytic reactions available for model training. Therefore, the 215 reactions discovered were combined with previously reported biocatalytic reactions associated with enzymes in aKGLib1 (Supplementary Information Section 12). Of the literature reactions, eight had been observed experimentally during our reaction discovery efforts and the extra 139 reactions were added to the dataset to create BioCatSet1 (Supplementary Table 3). To convert these substrate–enzyme pairs to inputs for machine learning models, reaction partners were divided into their individual components, comprising 119 substrates and 163 enzymes (Fig. 4a). Each substrate was converted to a SMILES string using the main protonation state at pH 7.5 (ref. 47), featurized using MORFEUS[46] and mapped to chemical space. To quantify the relationship between enzymes, the alignment scores were extracted from the SSN and converted to a normalized value (AS%) to capture the relationships as a quantitative matrix input.

With these metrics defined, we took steps towards building predictive models for navigating between chemical and sequence space. In a synthetic-chemistry-based endeavour, we constructed a substrate-to-enzyme recommendation system to enable the identification of new biocatalytic reactions with a given substrate. To achieve this, each substrate was mapped to chemical space and ten of its nearest

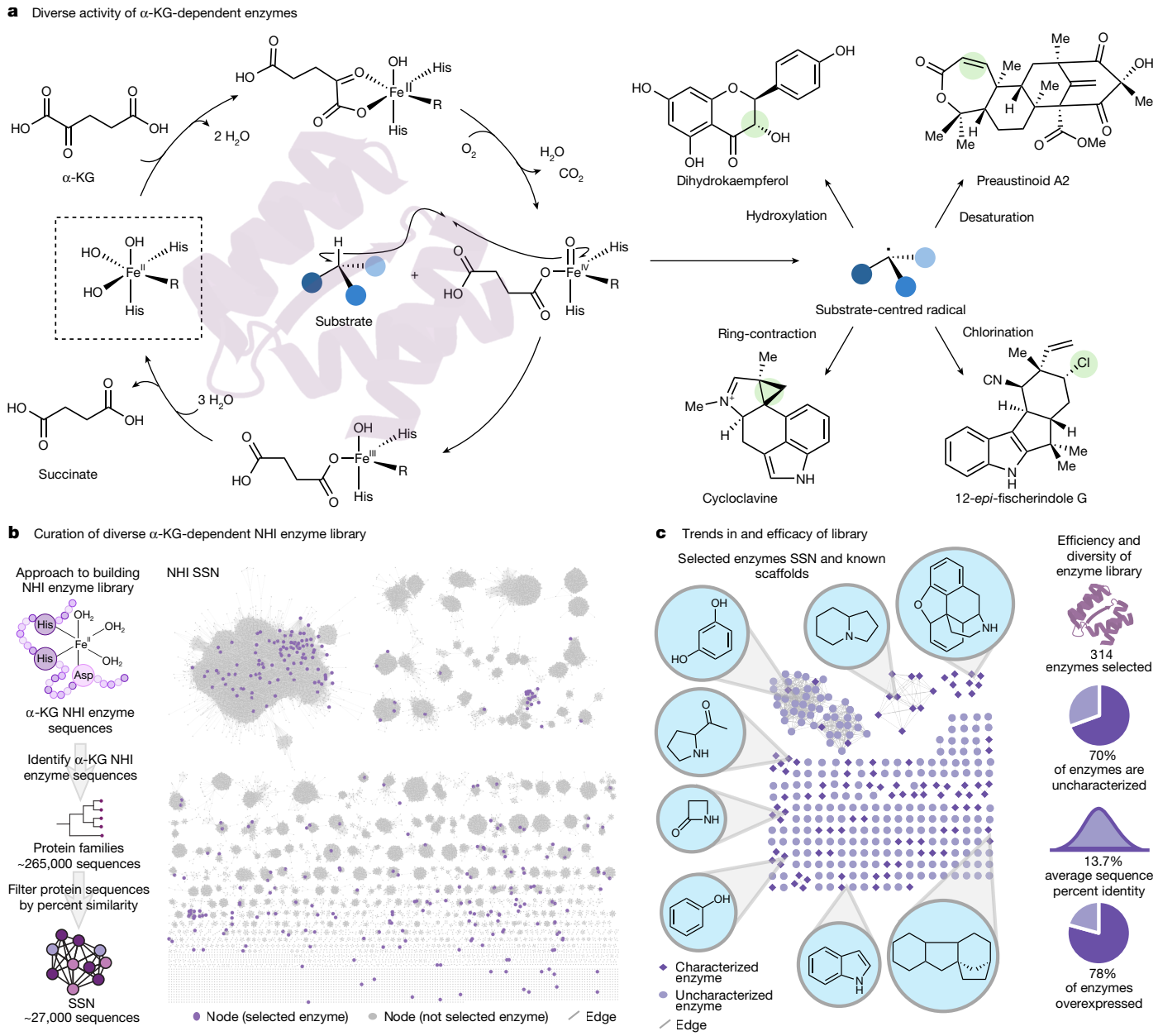

Fig. 2 | Rationale and curation of a diverse α-KG NHI enzyme library, aKGLib1. **a**, Abbreviated catalytic cycle and enzymatic transformations with α-KG-dependent C–H functionalization in natural product biosynthesis. In the active site of α-KG-dependent enzymes, iron is complexed by two histidine residues and either a carboxylate-containing residue (R = Asp/Glu) or an environmentally sourced halide (R = Ala/Gly). On α-KG binding and oxidation by atmospheric oxygen, the active iron(IV)-oxo species can initiate hydrogen atom abstraction from the substrate to yield the iron(III)-hydroxy species and a radical intermediate. This intermediate can undergo structural rearrangements before being terminated by rebound hydroxylation, carbocation formation or halogenation (functionalization by α-KG NHI enzymes in natural product biosynthesis shown in green) and generate succinate as a by-product. **b**, Workflow to curate a bioinformatics-guided α-KG-dependent NHI enzyme library (aKGLib1). The enzyme library was selected by collecting characterized of-interest enzyme sequences, which led to the inclusion of protein families IPR008775, IPR005123, IPR027443, IPR026992 and IPR044861. These families were used as a seed for the generation of a SSN (e-value = 5, UniRef90), which, after filtering, resulted in the network shown containing 27,005 protein sequences (alignment score = 50, organic full layout). 314 enzymes (purple) representing 1.16% of the total sequences (grey) were selected across 160 clusters, to generate a diverse enzyme library. **c**, Trends in substrates within clusters of a SSN and efficacy of aKGLib1. The sequences in the SSN at alignment score = 75 contain 94 enzymes that have previously been characterized (purple diamonds) and 220 sequences that are previously uncharacterized (lavender circles, 70% of total library). In clusters containing several characterized proteins, the known compatible common scaffold is highlighted. On performing a multisequence percent identity matrix, it was found that sequences only contained 13.7% shared identity, on average. On transformation and overexpression in *E. coli*, the presence of protein was investigated through gel electrophoresis, in which 78% of aKGLib1 showed soluble protein overexpressed at the expected molecular weight.

neighbours within the BioCatSet1 database were identified (Fig. 4b). The compatible enzymes for each neighbouring substrate were retrieved and the ten most similar enzymes within aKGLib1 were used to populate an output for the ranking model. The entirety of the output enzyme list, or subsets identified in decreasing order (k), were used to measure the precision@k, recall@k, enrichment@k and normalized discounted cumulative gain (nDCG)@k (Fig. 4b). precision@k measures the fraction of entries within the list that are known to be compatible with the

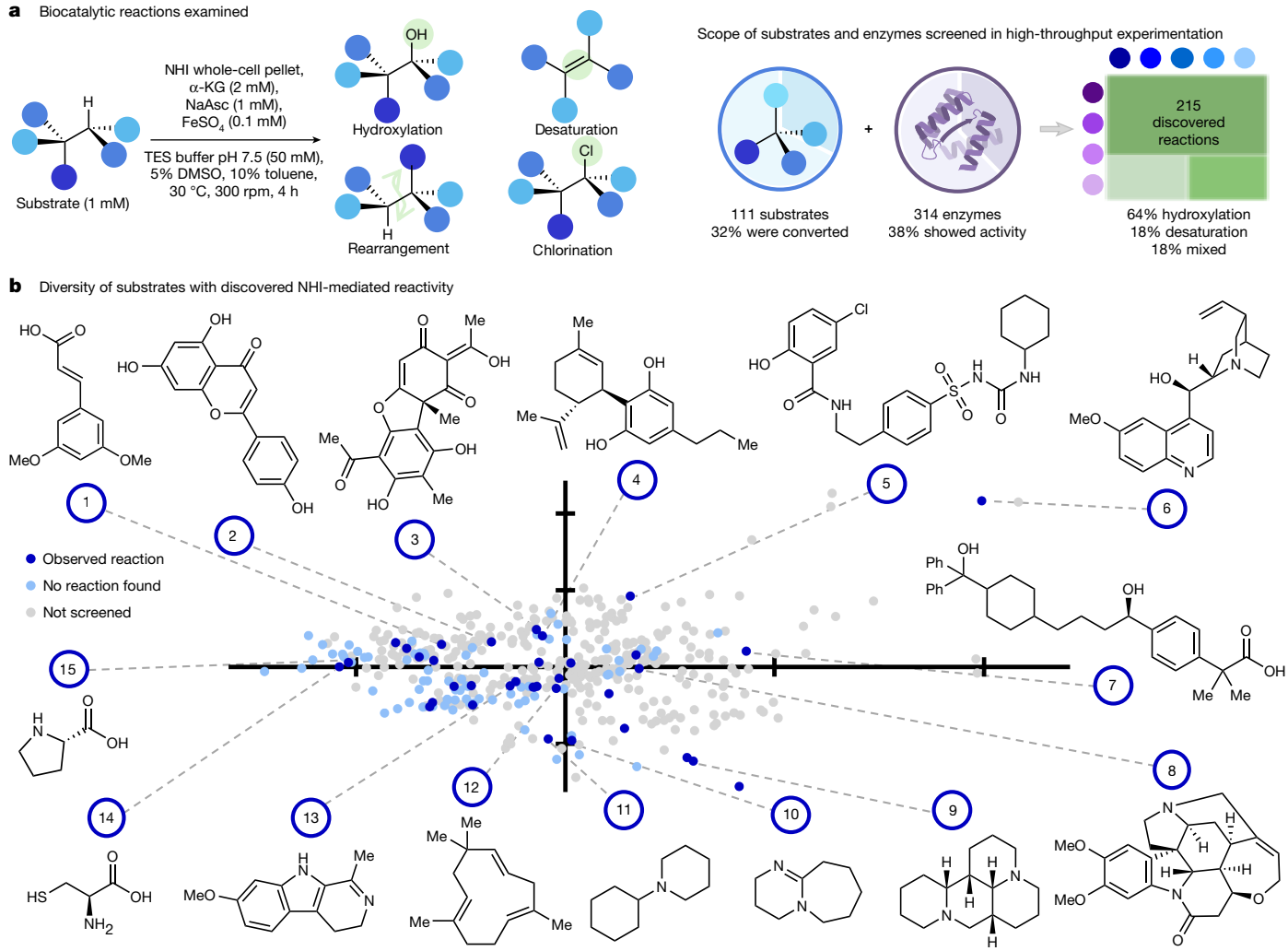

Fig. 3 | High-throughput reaction discovery workflow and diversity of biocatalytic reactions discovered. a, Biocatalytic reactions were investigated in 96-well plates, in which substrate (1 mM in DMSO) was added to a reaction mixture containing α-KG-dependent NHI enzymes in whole-cell pellet format (40% vol/vol of roughly 200 mg ml^{-1} 50 mM TES pH 7.5 suspensions), α-KG (2 mM in H_2O), sodium ascorbate (NaAsc, 1 mM in H_2O) and iron(II) sulfate heptahydrate ($FeSO_4$, 0.1 mM in H_2O) and then treated with toluene (10% vol/vol) to a final volume of 50 µl. Reactions were incubated at 30 °C while shaking at 300 rpm (15-mm orbital radius) for 4 h. Reactions were quenched, pooled and filtered to yield an analytical sample containing one enzyme with five substrates (0.100 µM) and all generated reaction products. Reactions were monitored by LC-MS using a 3-min reverse-phase method to identify unreacted substrate, hydroxylation, desaturation, rearrangement and chlorination products by mass (functionalization shown in green). Using this high-throughput reaction platform, 111 unique substrates were scrutinized against 314 NHI enzymes in triplicate. 32% of substrates underwent biocatalytic transformations and 38% of the enzymes showed activity with at least one substrate. This reaction platform generated 215 new compatible enzyme and substrate pairs. Discovered reactions are represented as a sunburst diagram, including hydroxylation (dark green, 64%), desaturation (light green, 18%) and mixed reaction outcomes (medium green, 14% hydroxylation/desaturation combination, 4% rearrangement). **b**, All substrates mapped in chemical space. Substrates with activity (dark blue, $n = 35$), substrates without any observed activity (light blue, $n = 76$) and further substrates (grey, $n = 335$) are represented in chemical space. This was generated by using the charge/protonation states of substrates at pH 7.5, calculating their MORFEUS descriptors, performing a principal component analysis and representing the substrates as values of PC1 and PC2. Part of the substrates with identified activity are shown in respect to their position in chemical space.

input and recall@k describes the fraction of entries compatible with the input that were populated within the prediction list. enrichment@k compares the precision with what would be achieved by randomly sampling BioCatSet1. nDCG@k reflects a weighted version of precision, for which entries ranked higher have a greater contribution to the total score. Ideally, at a low value of k, these metrics are high, signifying the curation of a streamlined rank list.

We anticipated that this dataset and approach could also be used to navigate from a given protein sequence to an area of chemical space to answer the question of which substrate a given enzyme might transform. Thus, we designed a complementary enzyme-to-substrate model. In this approach, each enzyme is compared with the members of aKGLib1 to identify its most similar sequences. The compatible substrates for each similar enzyme are retrieved and their nearest neighbours in chemical space are identified to generate a substrate rank list. In a similar fashion, precision@k, recall@k, enrichment@k and nDCG@k are calculated for the generated substrate prediction list.

After generating the BioCatSet1 dataset, designing the machine learning pipeline and establishing evaluation metrics, we trained an efficient model to navigate across substrate chemical and protein sequence space. A key step in this process was determining the most appropriate data-splitting strategy. For a rank-list-based task, the possible division of train–test data include substrate-oriented, enzyme-oriented or a simultaneous substrate/enzyme split (Fig. 4c). For the substrate-to-enzyme model, we implemented a 50/50 training/test split based on the substrate data and likewise performed a 50/50 training/test split based

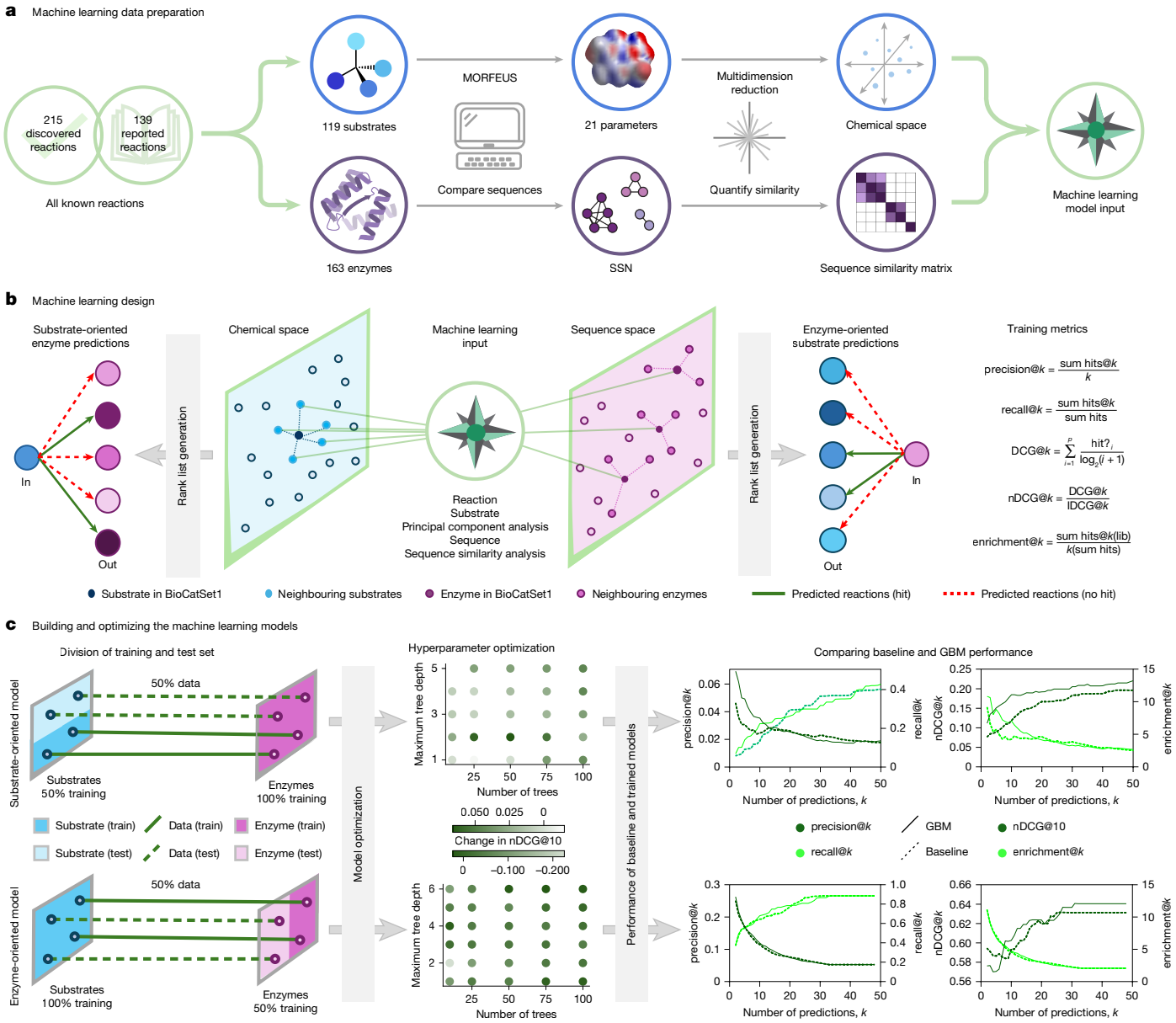

Fig. 4 | Machine learning approach, model building and output. a, The 215 newly discovered reactions were combined with all reported reactions (147 total, 139 not experimentally validated) for enzyme sequences in aKGLib1. 119 substrates, from discovered and literature reactions, were parameterized with MORFEUS before undergoing principal component analysis of the 21 features. Enzyme sequences were used to generate a SSN at alignment score = 5. Alignment scores were normalized as a fraction of the maximum percent identity to generate the alignment score percent (AS %). **b**, A machine learning model was designed to map a substrate (dark blue) or enzyme (dark purple) of interest, identify its nearest neighbours (light blue or light purple, respectively), identify their compatible enzyme sequences (dark purple) and substrates (dark blue) and populate a rank list with their nearest neighbours (light purple or light blue, respectively). Metrics to evaluate the rank lists are shown and optimization to populate the rank list with compatible enzymes and substrates at low k is represented. **c**, Division of train–test split in a substrate-to-enzyme (top) and enzyme-to-substrate (bottom) rank list task. For the substrate-oriented model, a substrate split was chosen to allow for the introduction of new substrates, without the high data cost of a dual split, and likewise in the enzyme-oriented model, an enzyme split was chosen. Training of a substrate-oriented gradient-boosted model identifies the optimal number of trees (50) and tree depth (2) and in the enzyme-oriented model, the optimal number of trees (100) and tree depth (6) were found. Machine learning models with the parameter optimization incorporated were evaluated against the baseline models to calculate precision, recall, nDCG and enrichment at various degrees of k. The GBM model is represented as a solid line and the baseline linear model is indicated by a dashed line. Precision and nDCG are shown in dark green and recall and enrichment are shown in light green.

on the enzyme data for the enzyme-to-substrate model. This unusual equal split between training and test sets was made to avoid potential biases caused by uneven dataset sizes.

As a baseline model, we built a simple substrate-to-enzyme prediction ranking formula that weighed the distance in chemical space over two dimensions and the distance in protein sequence space linearly. This model was able to enrich the predictions ranked in the top ten positions with enzymes compatible with each substrate >4-fold better than would be observed by randomly sampling aKGLib1 (Supplementary Table 8). To further minimize the experimental burden associated with enzyme screening, a model that rewards populating the enzyme prediction list with compatible enzymes at a low k was needed. After optimization, we found that distance over five dimensions of chemical space provided a more robust calculation for chemical similarity. To further improve performance, we trained a gradient-boosted decision tree ensemble model (GBM)[48] with YetiRank loss function[49].

This model offers an advantage with arbitrarily complex relationships and is relatively robust to overfitting when using proper hyperparameters[50]. The number of trees and their depth were optimized using a grid search procedure[50], yielding a model of 50 trees with a depth of 2 (Fig. 4c).

The performance of the substrate-to-enzyme GBM was compared with the baseline linear model. The GBM has a higher nDCG than the baseline model at all values of k. Furthermore, the top ten predicted enzymes are >7-fold more likely to be compatible with the input substrate than through random sampling of aKGLib1. At low values of k ($k < 20$), the GBM outperforms the baseline model in precision, indicating that the GBM is most well suited for the curation of a focused enzyme prediction list (Supplementary Table 8). Given this performance, we used a GBM to build the final workflow. Analysis of feature importance revealed that PC3, which mostly comprises two dipole moments and nucleofugality, has the greatest weight on the model, followed by PC1 (most heavily comprising dispersion descriptors, solvent-accessible surface area and volume) and the alignment score percent (Supplementary Figs. 30 and 31).

Following this model development, we trained an enzyme-to-substrate prediction rank model. The baseline model was constructed using the distance in five dimensions of chemical space and nearest-neighbours calculations were extended to include the training set of sequence space and entirety of chemical space in the algorithm. The hyperparameters of a GBM approach were optimized on the basis of nDCG at $k = 10$, yielding a model of 100 trees with a depth of 6. With these optimizations, the precision, recall, nDCG and enrichment were each measured at various rankings of k. Although minimal differences in these metrics were observed between the baseline and GBM models, these results highlight the generality of the GBM approach, even in challenging scenarios (Supplementary Tables 12 and 13).

CATNIP: a web app for prediction of biocatalytic reactions

With two machine learning models constructed, we created an interface to allow others to access predictions between α-KG NHI enzymes and small-molecule substrates. We created CATNIP (https://catnip.cheme.cmu.edu/), a web platform that allows scientists to interact directly with the substrate-to-enzyme and enzyme-to-substrate models. In the substrate-oriented model, users can input a chemical structure and receive a ranked list of aKGLib1 enzymes (and the corresponding sequences) potentially capable of transforming the targeted substrate (Supplementary Information Tutorial 1). Furthermore, users can gain insight into potential small-molecule substrates for NHI enzymes, using the enzyme-to-substrate model (Supplementary Information Tutorial 2). In this model, users can submit a protein sequence and receive a ranked list of small molecules beyond the scope of the training and test set that may be compatible. With the information provided by CATNIP, the user can execute a highly focused set of experiments to identify new biocatalytic reactivity. This strategy effectively derisks the implementation of biocatalysis in target-oriented synthesis by making use of machine learning.

The CATNIP substrate-to-enzyme workflow was tested with a selection of substrates, starting with the commercially available plant alkaloid sparteine (**16**, train set; Fig. 5a). Sparteine (**16**) was mapped to chemical space and CATNIP determined the ten nearest substrates within BioCatSet1, which includes highly decorated nitrogen heterocycles, specifically a piperidine, a bicyclic amidine, five indolizidines and three tetracyclic diamines. These neighbour substrates fed the machine learning model to generate a ranked list of enzymes, which contained four characterized enzyme sequences and six previously uncharacterized sequences, which were tested in reactions with sparteine (**16**). Seven of the ten reactions conducted resulted in a hydroxylation product as observed by LC-MS (Supplementary Fig. 33). The enzyme that produced the greatest amount of product on the analytical scale was used for a 50-mg-scale reaction from which hydroxylated product **17** was isolated in 35% isolated yield.

Similar success was achieved for more substrates tested in CATNIP. For example, matridine (**18**, train set), a synthetic precursor towards matrine natural products provided by Kerkovius et al.[51], was hydroxylated by seven of the top ten enzymes predicted. From a 50-mg-scale reaction, (12*S*)-hydroxymatridine (**19**) was isolated in 50% yield (Supplementary Fig. 34). Also, seven of the top ten enzymes predicted to transform 6-methyleneandrost-4-ene-3,17-dione (**20**, test set) led to productive reactions. On the preparative scale, **20** was converted in a 12% yield to the oxidative alkene cleavage product **21** (Supplementary Fig. 35). To the best of our knowledge, this is the first example of an α-KG NHI enzyme performing oxidative alkene cleavage of this type. Moreover, since the time that our machine learning models and predictive workflow were built, new reactions have been reported with this class of enzymes, providing extra test cases for CATNIP. For example, small molecules that Renata and colleagues experimentally determined as compatible substrates of enzymes within this library were in agreement with CATNIP substrate-to-enzyme outputs[52] (Supplementary Fig. 36).

The enzyme-to-substrate model was tested in a similar fashion (Fig. 5b). Using NHI123 from *Schizosaccharomyces pombe* (test set) as an input sequence, Clustal Omega was used to identify the ten most similar enzymes within aKGLib1. The substrates associated with these enzymes were retrieved, providing insight into the potential regions of chemical space compatible with these sequences. The ten best-ranked substrates, largely made up of monocyclic and bicyclic oxygen-containing molecules, were tested as substrates in reactions with NHI123. Four of these substrates were oxidized by NHI123. The top ranked prediction, substrate **22**, was transformed by NHI123 to a single product in 7% conversion (Supplementary Fig. 37). Similarly, the top ranked substrate for NHI177 from *Photorhabdus thracensis* (test set), humulene (**12**), was transformed by NHI177 to deliver a single oxidized product in 41% conversion (Supplementary Fig. 38). To test the accuracy of this model beyond enzymes within BioCatSet1, we submitted TqaL from *Streptomyces violaceusniger* (external validation) to CATNIP to identify the region of chemical space proposed to be compatible with the input enzyme sequence. The top 12 ranked substrates were subjected to analytical-scale reactions with TqaL, of which four were oxidized, including the second ranked substrate (**23**), providing an oxidized product in 42% conversion (Supplementary Fig. 39). Although this enzyme has no characterized activity, it is a homologue of a characterized enzyme that operates on similar amino acid substrates[53], reinforcing the performance of the model.

Conclusion

Overall, the development of this toolkit advances our ability to navigate between chemical and protein sequence space. Specifically, the curation of aKGLib1, a diverse NHI enzyme library comprising >300 wild-type proteins with low sequence identity and profiling of the biocatalytic activity of these enzymes against >100 small-molecule substrates led to the discovery of 215 new reactions. This dataset was combined with literature-reported reactions to make BioCatSet1, which was used to train two GBMs, generating substrate-to-enzyme and enzyme-to-substrate rank lists as outputs. With these models, we created CATNIP, an open-access web interface that enables streamlined biocatalytic reaction discovery. The power of these tools was demonstrated through expedited biocatalytic reaction discovery on substrates and enzymes outside the dataset. These reactions represent new connections between chemical and protein sequence space, creating opportunities for further exploration of the landscapes through substrate and protein engineering. We anticipate that this approach can be broadly applied to further enzyme families and reaction classes, offering a method to navigate between chemical and protein sequence

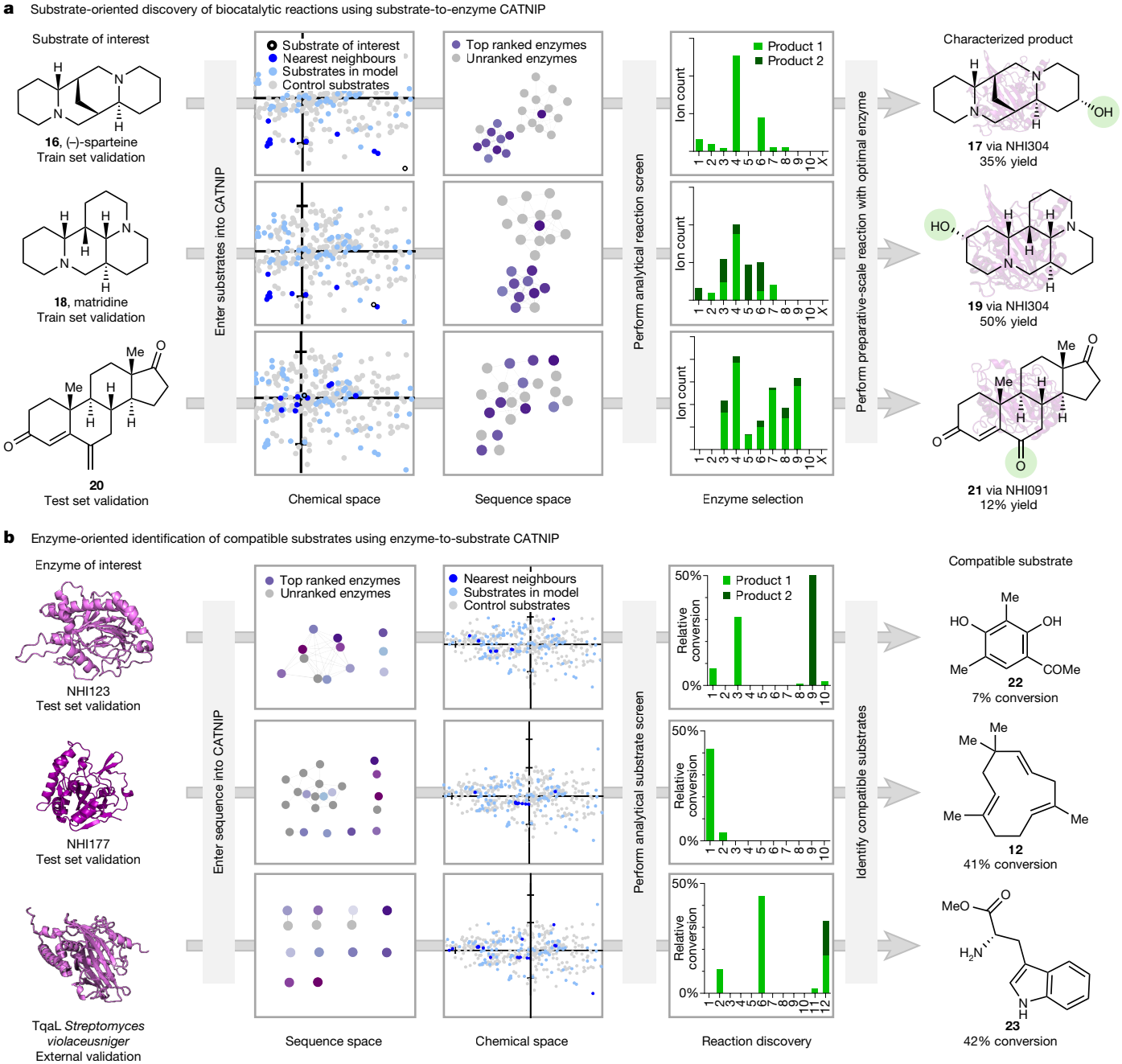

Fig. 5 | Use of the machine learning model for CATNIP. a, Demonstration of substrate-to-enzyme CATNIP with sparteine (**16**), matridine (**18**) and 6-methyleneandrost-4-ene-3,17-dione (**20**). In the chemical space map, the substrate of interest (open black circle), the nearest substrates over five dimensions (dark blue circle), unchosen substrates in BioCatSet1 (light blue circle) and substrates without known biocatalytic activity (grey circle) are shown. The sequence space shows all enzymes in the cluster (SSN at alignment score = 75) with predicted compatible enzymes ($k = 10$), with rank shown in decreasing shades of purple. Enzymes not predicted in the top ten sequences are represented as grey nodes. The top ten predicted enzyme sequences were prepared in whole-cell *E. coli* and examined for relative product formation in triplicate. The *x*-axis contains the enzyme prediction rank, for which X = no enzyme control. The *y*-axis shows the average relative extracted ion count ($n = 3$). Several products are represented with various shades of green. The enzyme generating the most product was then produced (1 -l cultures in Terrific Broth) and used in 50-mg-scale biocatalytic reactions as clarified cell lysate. Oxidation products were isolated and characterized for the three substrates of interest, providing (4*S*)-hydroxysparteine (**17**), (12*S*)-hydroxymatridine (**19**) and androst-4-ene-3,6,17-trione (**21**) in 35%, 50% and 12% isolated yields, respectively. **b**, Demonstration of the enzyme-to-substrate CATNIP model with NHI123, NHI177 and TqaL. Each enzyme was mapped to sequence space, which shows all enzymes in the cluster (SSN at alignment score = 75), with the ten most similar enzymes shown in decreasing shades of purple. Enzymes not predicted in the top ten sequences are represented as grey nodes. The predicted compatible substrates are identified (dark blue) and mapped to chemical space among all substrates in BioCatSet1 (light blue) and substrates outside the dataset (grey). The best-ranked substrates were tested with the enzyme of interest in triplicate and the relative product conversion was measured. The *x*-axis shows the rank of the small molecule substrate in decreasing order. The *y*-axis shows the average normalized relative conversion, as compared with the empty vector control of each sample ($n = 3$). The structure for the best-ranked substrate for each enzyme is shown as **22**, **12** and **23**, respectively.

space. Furthermore, this innovation effectively derisks the application of biocatalysts in organic synthesis.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-025-09519-5.

1. Li, J., Amatuni, A. & Renata, H. Recent advances in the chemoenzymatic synthesis of bioactive natural products. *Curr. Opin. Chem. Biol.* **55**, 111–118 (2020).
2. Romero, E. et al. Enzymatic late-stage modifications: better late than never. *Angew. Chem. Int. Ed.* **60**, 16824–16855 (2021).
3. Bayer, T., Wu, S., Snajdrova, R., Baldenius, K. & Bornscheuer, U. T. An update: enzymatic synthesis for industrial applications. *Angew. Chem. Int. Ed.* **64**, e202505976 (2025).
4. Marshall, J. R., Mangas-Sanchez, J. & Turner, N. J. Expanding the synthetic scope of biocatalysis by enzyme discovery and protein engineering. *Tetrahedron* **82**, 131926 (2021).
5. Yang, J., Li, F.-Z. & Arnold, F. H. Opportunities and challenges for machine learning-assisted enzyme engineering. *ACS Cent. Sci.* **10**, 226–241 (2024).
6. Bell, E. L. et al. Biocatalysis. *Nat. Rev. Methods Primers* **1**, 46 (2021).
7. Buller, R. et al. From nature to industry: harnessing enzymes for biocatalysis. *Science* **382**, eadh8615 (2023).
8. Garzón-Posse, F., Becerra-Figueroa, L., Hernández-Arias, J. & Gamba-Sánchez, D. Whole cells as biocatalysts in organic transformations. *Molecules* **23**, 1265 (2018).
9. Tibrewal, N. & Tang, Y. Biocatalysts for natural product biosynthesis. *Annu. Rev. Chem. Biomol. Eng.* **5**, 347–366 (2014).
10. Roiban, G.-D. et al. Development of an enzymatic process for the production of (*R*)-2-butyl-2-ethyloxirane. *Org. Process Res. Dev.* **21**, 1302–1310 (2017).
11. Arnold, F. H. Directed evolution: bringing new chemistry to life. *Angew. Chem. Int. Ed.* **57**, 4143–4148 (2018).
12. Tobin, P. H., Richards, D. H., Callender, R. A. & Wilson, C. J. Protein engineering: a new frontier for biological therapeutics. *Curr. Drug. Metab.* **15**, 743–756 (2014).
13. Novick, S. J. et al. Engineering an amine transaminase for the efficient production of a chiral sacubitril precursor. *ACS Catal.* **11**, 3762–3770 (2021).
14. Lovelock, S. L. et al. The road to fully programmable protein catalysis. *Nature* **606**, 49–58 (2022).
15. Yu, T. et al. Enzyme function prediction using contrastive learning. *Science* **379**, 1358–1363 (2023).
16. Hon, J. et al. EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities. *Nucleic Acids Res.* **48**, W104–W109 (2020).
17. Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.* **5**, e1000605 (2009).
18. Robertson, D. E. et al. Exploring nitrilase sequence space for enantioselective catalysis. *Appl. Environ. Microbiol.* **70**, 2429–2436 (2004).
19. Wahler, D., Badalassi, F., Crotti, P. & Reymond, J.-L. Enzyme fingerprints by fluorogenic and chromogenic substrate arrays. *Angew. Chem. Int. Ed.* **40**, 4457–4460 (2001).
20. Finnigan, W., Hepworth, L. J., Flitsch, S. L. & Turner, N. J. RetroBioCat as a computer-aided synthesis planning tool for biocatalytic reactions and cascades. *Nat. Catal.* **4**, 98–104 (2021).
21. Fansher, D. J., Besna, J. N., Fendri, A. & Pelletier, J. N. Choose your own adventure: a comprehensive database of reactions catalyzed by cytochrome P450 BM3 variants. *ACS Catal.* **14**, 5560–5592 (2024).
22. Ma, E. J. et al. Machine-directed evolution of an imine reductase for activity and stereoselectivity. *ACS Catal.* **11**, 12433–12445 (2021).
23. Ao, Y.-F. et al. Structure- and data-driven protein engineering of transaminases for improving activity and stereoselectivity. *Angew. Chem. Int. Ed.* **62**, e202301660 (2023).
24. Supekar, S. et al. A machine learning-guided approach to navigate the substrate activity scope of galactose oxidase: application in the conversion of pharmaceutically relevant bulky secondary alcohols. *ACS Catal.* **14**, 17233–17243 (2024).
25. King, B. R., Sumida, K. H., Caruso, J. L., Baker, D. & Zalatan, J. G. Computational stabilization of a non-heme iron enzyme enables efficient evolution of new function. *Angew. Chem. Int. Ed.* **64**, e202414705 (2025).
26. Mou, Z. et al. Machine learning-based prediction of enzyme substrate scope: application to bacterial nitrilases. *Proteins* **89**, 336–347 (2021).
27. Yang, M. et al. Functional and informatics analysis enables glycosyltransferase activity prediction. *Nat. Chem. Biol.* **14**, 1109–1117 (2018).
28. Kroll, A., Ranjan, S., Engqvist, M. K. M. & Lercher, M. J. A general model to predict small molecule substrates of enzymes based on machine and deep learning. *Nat. Commun.* **14**, 2787 (2023).
29. Goldman, S., Das, R., Yang, K. K. & Coley, C. W. Machine learning modeling of family wide enzyme-substrate specificity screens. *PLoS Comput. Biol.* **18**, e1009853 (2022).
30. Wang, X., Quinn, D., Moody, T. S. & Huang, M. ALDELE: all-purpose deep learning toolkits for predicting the biocatalytic activities of enzymes. *J. Chem. Inf. Model.* **64**, 3123–3139 (2024).
31. Busch, F., Brummund, J., Calderini, E., Schürmann, M. & Kourist, R. Cofactor generation cascade for α-ketoglutarate and Fe(II)-dependent dioxygenases. *ACS Sustain. Chem. Eng.* **8**, 8604–8612 (2020).
32. Zwick, C. R. & Renata, H. Harnessing the biocatalytic potential of iron- and α-ketoglutarate-dependent dioxygenases in natural product total synthesis. *Nat. Prod. Rep.* **37**, 1065–1079 (2020).
33. Gao, S. S., Naowarojna, N., Cheng, R., Liu, X. & Liu, P. Recent examples of α-ketoglutarate-dependent mononuclear non-haem iron enzymes in natural product biosyntheses. *Nat. Prod. Rep.* **35**, 792–837 (2018).
34. Hausinger, R. P. Fe(II)/α-ketoglutarate-dependent hydroxylases and related enzymes. *Crit. Rev. Biochem. Mol. Biol.* **39**, 21–68 (2004).
35. McLean, K. J., Luciakova, D., Belcher, J., Tee, K. L. & Munro, A. W. Biological diversity of cytochrome P450 redox partner systems. *Adv. Exp. Med. Biol.* **851**, 299–317 (2015).
36. Schofield, C. J. & Zhang, Z. Structural and mechanistic studies on 2-oxoglutarate-dependent oxygenases and related enzymes. *Curr. Opin. Struct. Biol.* **9**, 722–731 (1999).
37. Seide, S. et al. From enzyme to preparative cascade reactions with immobilized enzymes: tuning Fe(II)/α-ketoglutarate-dependent lysine hydroxylases for application in biotransformations. *Catalysts* **12**, 354 (2022).
38. Hegg, E. L. & Que, L. Jr The 2-His-1-carboxylate facial triad — an emerging structural motif in mononuclear non-heme iron(II) enzymes. *Eur. J. Biochem.* **250**, 625–629 (1997).
39. Zallot, R., Oberg, N. & Gerlt, J. A. The EFI web resource for genomic enzymology tools: leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. *Biochemistry* **58**, 4169–4182 (2019).
40. Fisher, B. F., Snodgrass, H. M., Jones, K. A., Andorfer, M. C. & Lewis, J. C. Site-selective C–H halogenation using flavin-dependent halogenases identified via family-wide activity profiling. *ACS Cent. Sci.* **5**, 1844–1856 (2019).
41. Atkinson, H. J., Morris, J. H., Ferrin, T. E. & Babbitt, P. C. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One* **4**, e4345 (2009).
42. Copp, J. N., Akiva, E., Babbitt, P. C. & Tokuriki, N. Revealing unexplored sequence-function space using sequence similarity networks. *Biochemistry* **57**, 4651–4662 (2018).
43. Pyser, J. B. et al. Stereodivergent, chemoenzymatic synthesis of azaphilone natural products. *J. Am. Chem. Soc.* **141**, 18551–18559 (2019).
44. Lima, S. T. et al. A widely distributed biosynthetic cassette is responsible for diverse plant side chain cross-linked cyclopeptides. *Angew. Chem. Int. Ed.* **62**, e202218082 (2023).
45. Ju, S. et al. A biocatalytic platform for asymmetric alkylation of α-keto acids by mining and engineering of methyltransferases. *Nat. Commun.* **14**, 5704 (2023).
46. Jacot-Descombes, L., Turcani, L. & Jorner, K. morfeus (computer software). https://github.com/digital-chemistry-laboratory/morfeus (accessed 29 August 2025).
47. Ropp, P. J., Kaminsky, J. C., Yablonski, S. & Durrant, J. D. Dimorphite-DL: an open-source program for enumerating the ionization states of drug-like small molecules. *J. Cheminform.* **11**, 14 (2019).
48. Hastie, T., Tibshirani, R. & Friedman, J. in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 605–624 (Springer, 2009).
49. Lyzhin, I., Ustimenko, A., Gulin, A. & Prokhorenkova, L. Which tricks are important for learning to rank? *Proc. 40th Intl Conf. Machine Learning* (ICML 2023), *PMLR* **202**, 23264–23278 (2023).
50. Bentéjac, C., Csörgő, A. & Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **54**, 1937–1967 (2021).
51. Kerkovius, J. K. et al. A pyridine dearomatization approach to the matrine-type lupin alkaloids. *J. Am. Chem. Soc.* **144**, 15938–15943 (2022).
52. Xu, H., Zhao, J. & Renata, H. Discovery, characterization and synthetic application of a promiscuous nonheme iron biocatalyst with dual hydroxylase/desaturase activity. *Angew. Chem. Int. Ed.* **63**, e202409143 (2024).
53. Bunno, R., Awakawa, T., Mori, T. & Abe, I. Aziridine formation by a Fe^{II}/α-ketoglutarate dependent oxygenase and 2-aminoisobutyrate biosynthesis in fungi. *Angew. Chem. Int. Ed.* **60**, 15827–15831 (2021).

Data availability

All of the experiments discussed are provided in the Supplementary Information. The data are available at https://huggingface.co/gomes-group/catnip.

Code availability

All of the code is available at https://doi.org/10.5281/zenodo.16779318 (ref. 54).

54. Paton, A. E. et al. Connecting chemical and protein sequence space to predict biocatalytic reactions (v0.1). *Zenodo* https://doi.org/10.5281/zenodo.16779318 (2024).

Acknowledgements This research was supported by funds from the Life Sciences Institute and the Department of Chemistry at the University of Michigan. Part of the enzyme library was built in collaboration with the Joint Genome Institute (JGI) DNA Synthesis Program. Efforts to profile the reactivity of the enzyme library were supported by the Novartis Global Scholars Program. Aspects of the data science components of this project were supported by the Camille Dreyfus Teacher-Scholar Award (TC-20-061) and the MBioFAR Program at the University of Michigan (A.R.H.N.). This work is part of the National Science Foundation Center for Chemoenzymatic Synthesis (2221346, A.R.H.N. and G.G.). G.G. thanks Carnegie Mellon University and the departments of chemistry and chemical engineering for the start-up support. We thank S. Reisman (California Institute of Technology), K. Brummond (University of Pittsburgh), J. Schomaker (University of Wisconsin) and J. Morgan (University of North Carolina Wilmington) for their generous contribution of complex molecules and K. V. Brown (University of Michigan) and R. Snajdrova and E. Tassano (Novartis Institutes for Biomedical Research, Inc.) for helpful discussions.

Author contributions A.R.H.N., J.C.P. and A.E.P. conceptualized the high-throughput generation of connections between chemical space and protein sequence space. J.C.P. performed bioinformatic investigations, designed the enzyme library and performed SDS-PAGE analysis. A.E.P. developed the experimental high-throughput reaction platform, performed biocatalytic reactions, analysed LC-MS data, proposed the concept for chemical and sequence space navigation, designed the baseline metric for enzyme ranking and executed part of the preparative-scale biocatalytic chemistry. T.R. generated initial MORFEUS descriptors. G.G. and D.A.B. designed and implemented the machine learning pipeline, including chemical space analysis, candidate generation and reranking procedures, as well as developed the CATNIP platform. N.I.C. assisted in high-throughput reaction discovery and performed part of the preparative-scale biocatalytic chemistry. A.R.H.N. and G.G. secured funding and oversaw the project. A.E.P., D.A.B., G.G. and A.R.H.N. wrote the manuscript, with input from all authors.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-025-09519-5.

Correspondence and requests for materials should be addressed to Gabe Gomes or Alison R. H. Narayan.

Peer review information *Nature* thanks Uwe Bornscheuer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.