

Structure-Based Design of Small Imine Reductase Panels for Target Substrates

Yuqi Yu,* Arnau Rué Casamajo, William Finnigan, Christian Schnepel, Rhys Barker, Charlotte Morrill, Rachel S. Heath, Leonardo De Maria, Nicholas J. Turner,* and Nigel S. Scrutton*



Cite This: *ACS Catal.* 2023, 13, 12310–12321



Read Online

ACCESS |

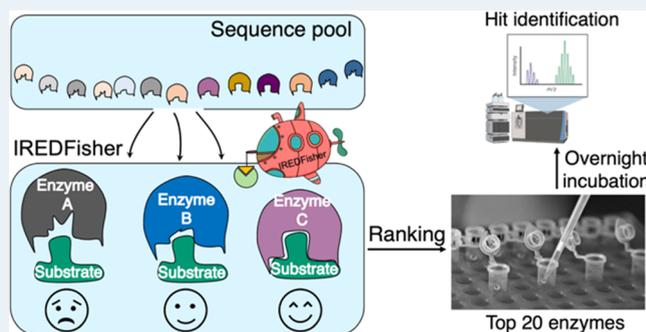
Metrics & More

Article Recommendations

Supporting Information

ABSTRACT: Biocatalysis is important in the discovery, development, and manufacture of pharmaceuticals. However, the identification of enzymes for target transformations of interest requires major screening efforts. Here, we report a structure-based computational workflow to prioritize protein sequences by a score based on predicted activities on substrates, thereby reducing a resource-intensive laboratory-based biocatalyst screening. We selected imine reductases (IREDs) as a class of biocatalysts to illustrate the application of the computational workflow termed IREDFisher. Validation by using published data showed that IREDFisher can retrieve the best enzymes and increase the hit rate by identifying the top 20 ranked sequences. The power of IREDFisher is confirmed by computationally screening 1400 sequences for chosen reductive amination reactions with different levels of complexity. Highly active IREDs were identified by only testing 20 samples in vitro. Our speed test shows that it only takes 90 min to rank 85 sequences from user input and 30 min for the established IREDFisher database containing 591 IRED sequences. IREDFisher is available as a user-friendly web interface (<https://enzymeevolver.com/IREDFisher>). IREDFisher enables the rapid discovery of IREDs for applications in synthesis and directed evolution studies, with minimal time and resource expenditure. Future use of the workflow with other enzyme families could be implemented following the modification of the workflow scoring function.

KEYWORDS: biocatalysis, enzyme screening, imine reductase, computational workflow, structural modeling



INTRODUCTION

Enzymatic synthesis has gained growing attention in recent years due to the inherent advantages of benign reaction conditions along with exquisite chemo-, stereo-, and regioselectivity.^{1–5} However, finding the optimal biocatalyst to effect the transformation of a substrate of interest is often challenging and time consuming. There are two strategies commonly used in enzyme discovery (Figure 1, left panel). One is to manually select the enzymes that have been reported to be active toward similar substrate(s) in the literature. However, this approach restricts the diversity of accessible enzymes and may result in potentially active enzymes being missed. Consequently, the alternative strategy of searching for homologous enzymes from databases is increasingly applied. However, this approach typically generates a very large number of sequences, the majority of which have not been characterized experimentally.^{6,7} Screening all of these sequences is time consuming, requiring the extensive use of analytical equipment and other methods for high-throughput screening.

Computational tools have shown great power in the development and application of biocatalysts including synthetic pathway design,^{8–10} enzyme engineering,^{11–14} and de novo design.^{15–18} However, in silico pipelines that prioritize

sequences and thus reduce the necessary screening throughput in the laboratory are still underdeveloped. The current structure-based pipelines are either struggling in dealing with large numbers of sequences and are thus mainly used for engineering an initially known active enzyme^{19,20} or require a detailed description of the catalytic geometry based on a known ligand-binding motif.^{21–23} Sequence-based pipelines^{24–38} can realize high throughput but they generally ignore substrate variations (a key consideration in establishing biocatalytic panels), leading to poor activity predictions of enzymes in relation to target substrates. Standalone methods such as homology modeling and molecular docking have been established over many years but are generally used individually rather than in automated combination, which limits throughput in biocatalyst screening programs. Integration of these approaches into an automated pipeline for biocatalyst

Received: May 19, 2023

Revised: August 20, 2023

Published: September 5, 2023



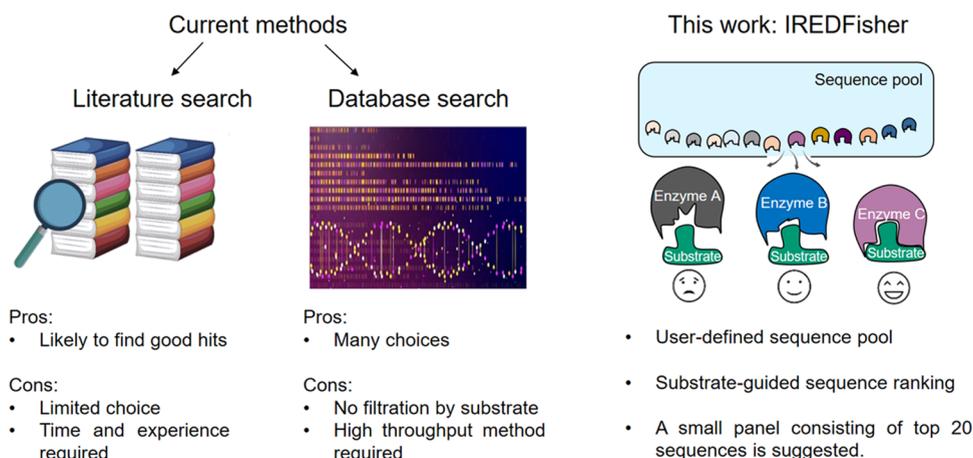


Figure 1. Solutions of enzyme discovery for targeted substrate(s). Left: two common methods for finding hit enzymes for targeted substrate(s). Right: a schematic diagram of IREDFisher developed in this work. Input sequences are filtered by IREDFisher with the aid of structural modeling and molecular docking, which allows the user to obtain a small panel consisting of prioritized sequences.

discovery is urgently needed to rapidly model enzyme–substrate complexes and generate rank ordering of likely active sequences. Only a few structure-based computational pipelines^{38–41} have been published. However, none of these have been automated to enable easy access to laboratory-based scientists. An “easy-to-use” *in silico* web application is required for experimentalists who need to conduct efficient and rapid screens to identify active enzymes but are struggling to use more specialized computational tools. Consequently, an “easy-to-use” *in silico* web-based application will streamline experimental screening and thereby reduce the time and resources required to identify biocatalysts for use.

Here, we report the development and experimental validation of an *in silico* pipeline, available also via a web interface, termed IREDFisher (available for use at <https://enzymeevolver.com/IREDFisher>). IREDFisher is a structure-based web server that allows computer-aided enzyme screening with substrate(s) of interest and recommends a small (20 enzyme) panel of IREs for rapid biocatalytic exploration by the user (Figure 2). We selected imine reductases (IREs) as a model enzyme in view of their recent emergence as an enzyme class of interest for the synthesis of chiral amines. Chiral amines are prevalent motifs that feature in about 40% of pharmaceuticals.^{42,43} IREs catalyze enantioselective imine reduction and can also perform the reductive amination of ketones or aldehydes^{44–46} and are thus of major interest to the academic and industrial biocatalysis communities. Since their discovery in 2010,⁴⁷ a vast number of IRE sequences have been published together with a wealth of activity data for different substrates.^{48–58} Currently, IREDFisher only applies to IREs as the scoring function is modified specifically for these enzymes.

RESULTS

IREDFisher Workflow. Having the initial sequences and a substrate structure as inputs, the IREDFisher workflow commences with sequence preprocessing, followed by structural modeling. Then, the provided substrate is docked into the modeled enzyme structures, while the protein–substrate complexes are scored by considering predicted binding affinity and catalytic geometry in the final step (Figure 2). Specifically, in step (1), each input sequence is aligned to individual sequences in the protein data bank (PDB). The

input sequence that has the highest sequence identity to known IREs in the PDB is considered as a putative IRE and retained for the next step. In step (2), all three-dimensional structures of the resulting sequences from step (1) are generated by MODELLER⁵⁹ followed by an optimization by the default function in MODELLER,⁵⁹ which consists of spatial restraints on distance and dihedral angles obtained from the target structure and CHARMM energy terms enforcing proper stereochemistry.⁶⁰ All models are evaluated based on the Ramachandran plot, and only those that have over 90% of the residues in favorable/coregions are defined as qualified models and subsequently used for substrate docking. In step (3), the substrate is docked into the active site of each qualified model. Various poses of the substrate in the active site of the enzymes are generated. In step (4), the distance between NADPH C4 and N atom of the imine substrate is calculated for all poses. The pose with ~ 4 Å is selected as the best pose to realize hydride transfer in a catalytic process (see pose selection in Figure 2 and details in the Methods section). In step (5), the number of acidic residues N_{acidic} , basic residues N_{basic} , and histidine residues N_{His} are calculated and used to modify the Autodock Vina⁶¹ scoring function considering they play important roles in the stabilization of iminium intermediate formation and proton transfer (see details in the Methods section). The modified scoring function Refined score = $4.0 \times \text{Vina score} + 1.0 \times N_{\text{acidic}} - 9.0 \times N_{\text{His}} + 9.0 \times N_{\text{basic}}$ was used to finally rank all input sequences. The top 20 IRE sequences are then selected for *in vitro* screening.

IREDFisher Validation Using Published Screening Data. The screening data from three established IRE panels were used to test the IREDFisher workflow: Roiban et al. panel⁵⁶ (85 sequences) for substrates 1–9, France et al. panel⁵² (45 sequences) for substrates 10–17, and Montgomery et al. IRE panel⁵³ (93 sequences) for 18–24 (Figure 3). Each panel and the corresponding tested substrates were fed into IREDFisher to obtain the recommended 20 sequences. As control, 20 sequences were randomly selected from each corresponding panel. Hit rates using a conversion cutoff of 2% (hit rate of $\geq 2\%$ conversion) and 50% (hit rate of $\geq 50\%$ conversion) were calculated and compared between the two methods (Figure 4a,b and Table S1). The likelihood of retrieving the best hit in the 20 selected sequences using the two methods was also calculated (Figure 4c and Table S2). In

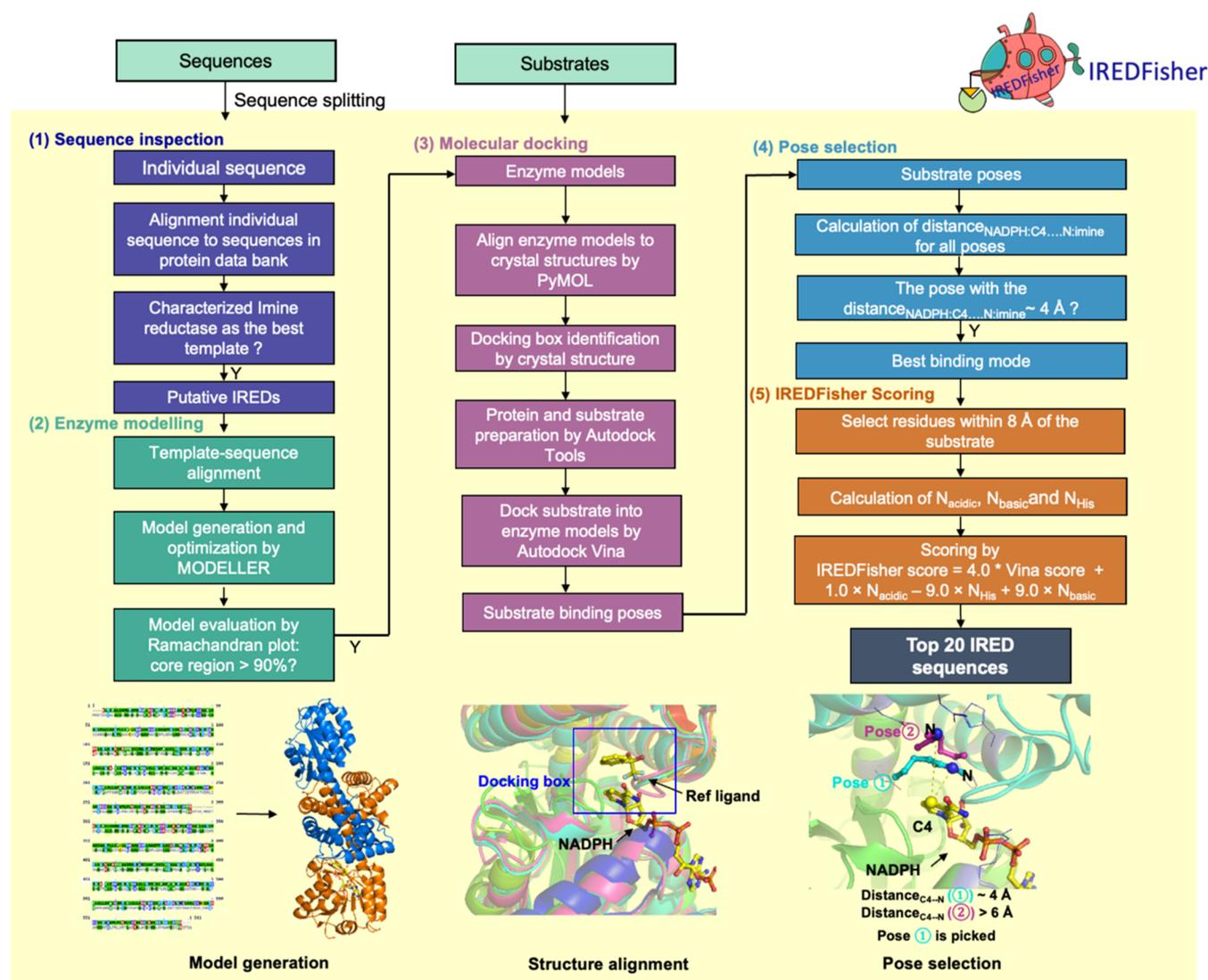


Figure 2. Flow chart of the four-step IREDFisher workflow. The input sequences are preprocessed in step (1) to remove non-IREd-homologs from the screening list. The three-dimensional structures of the remaining sequences are modeled in step (2). In step (3), the structures obtained from step (2) are aligned to a cofactor- and ligand-bound IRED crystal structure to locate the active site and cofactor binding site for each modeled structure. NADPH is placed into individual modeled structures by taking the coordinates from the crystal structure. An intermediate imine structure formed by amine and ketone substrates is then docked into the active site of each structure. In step (4), the best pose for the substrate within the enzyme is selected by the distance between two key atoms of the cofactor and substrate. In crystal structures of IREds, this distance is around 4 Å. In step (5), the number of acidic residues N_{acidic} , the number of basic residues N_{basic} , and the presence of histidine residues $N_{\text{His}} = 0$ or 1 are calculated for residues within 8 Å of the substrate. Enzymes are ranked by a modified scoring function $\text{IREDFisher score} = 4.0 \times \text{Vina score} + 1.0 \times N_{\text{acidic}} - 9.0 \times N_{\text{His}} + 9.0 \times N_{\text{basic}}$ (see details described in the [Methods](#) section).

terms of “hit rate of $\geq 2\%$ conversion”, IREDFisher is comparable to random selection, whereas it outperforms the latter in finding hits with conversions over 50% (Figure 4b). This shows the capability of IREDFisher in finding competent hits for downstream optimization. Furthermore, IREDFisher retrieved the best hit to the top 20 sequences with a likelihood of 100%. By contrast, the likelihood of picking the best hit by random selection is no more than 50%. In summary, the IREDFisher workflow is not only able to improve the hit rate but also retrieves the best enzyme(s) from the whole IRED panel in the top 20 selected sequences. This emphasizes the effectiveness of IREDFisher in finding IRED hits, reducing screening efforts by 50–75%.

Experimental Validation of Predictions Made by IREDFisher. Next, we challenged the IREDFisher workflow

by aiming to select potential IRED hits for subsequent validation by experimental screening. We chose the following reductive amination reactions as our targets, namely, **18** and **25–28** (Figure 5a), which represent increasingly demanding conversions for imine reductases. Reactions corresponding to products **18** and **25** are common starting points for screening IREds, whereas products **26** and **27** are important building blocks in the pharmaceutical industry and proved to be challenging targets in a previous IRED study.⁴⁸ Synthesis of **28** is particularly challenging due to the negative charge of the carboxylate group in the substrate. To date, no IREds have been reported to accept 4-formylbenzoic acid as a substrate. Also, we note that 4-formylbenzoic acid **28** is a building block for drugs such as procarbazine (chemotherapy), imatinib (chemotherapy), mocetinostat, etinostat, tucidinostat (HDAC

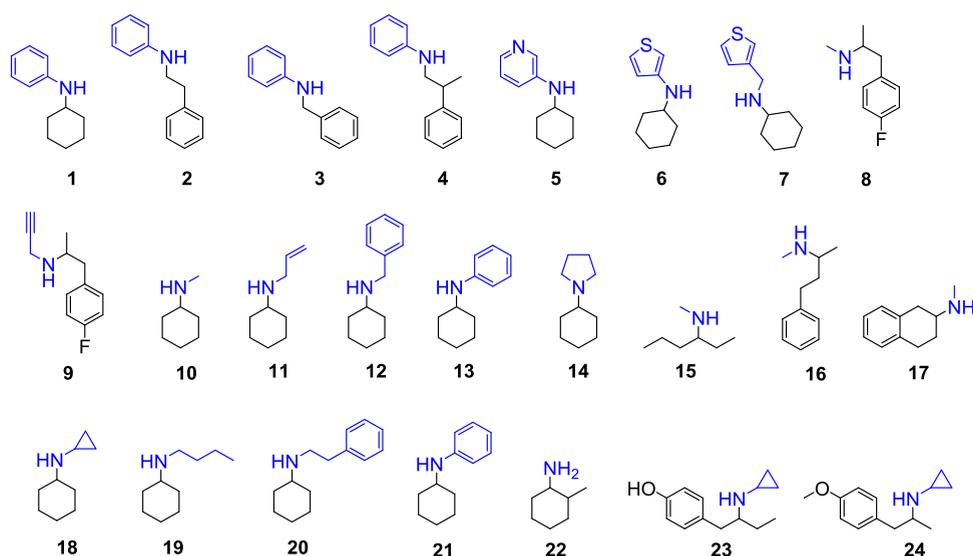


Figure 3. Tests of reductive amination reactions based on published data. Reaction products of IRED-catalyzed reactions are shown with moieties originating from the amine nucleophile colored in blue and those from the ketone component displayed in black. Compounds 1–9 are from ref 56, compounds 10–17 are from ref 52, and compounds 18–24 are from ref 53.

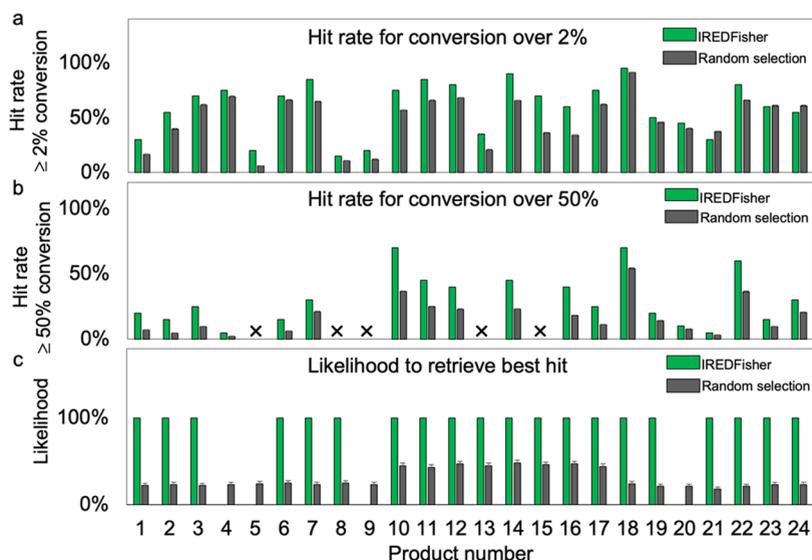


Figure 4. Comparison of hit rate and best hit retrieval in 20 sequences selected by IREDFisher and by random selection. (a) Hit rate using a conversion cutoff of 2% calculated by the equation: Hit rate of $\geq 2\%$ conversion = $\frac{\text{number of enzymes } (\geq 2\% \text{ conversion})}{20} \times 100\%$. (b) Hit rate using a conversion cutoff of 50% calculated by the equation: Hit rate of $\geq 50\%$ conversion = $\frac{\text{number of enzymes } (\geq 50\% \text{ conversion})}{20} \times 100\%$. Symbol \times indicates that no hits with conversion over 50% were found in the in vitro screening of the corresponding panels. (c) Likelihood of retrieving the best hit from the whole panel in the 20 selected sequences using IREDFisher. The error bars for the random selection were calculated by the standard error of the mean (SEM) based on 95% confidence levels.

inhibitors), revefenacin (COPD treatment), bavisant (ADHD treatment), and fedovapagon (nocturia treatment) (Figure S1). Inclusion of 28 was used to explore whether the IREDFisher workflow could identify biocatalysts for novel reductive amination reactions.

For sequence input, we used 193 well-expressed enzymes from our in-house panel along with 33 new sequences found by IREDFisher that are also expressed well. To find new active enzymes toward selected targets, we collected uncharacterized IRED homologs from public databases (NCBI and UniProt) and IREDS from established panels (Robian et al.,⁵⁶ France et al.,⁵² Wetzl et al.,⁵⁴ and Yao et al.⁵¹), generating 1239

sequences in total (Figure 5a, lower panel). Then, the first round of the IREDFisher ranking was run for the collected sequences based on the targeted product to prioritize successful gene synthesis and protein expression. The top 20 enzyme sequences for targets 18 and 25–30 were selected in silico. Then, 60 sequences were identified after removing duplicates from which genes were synthesized and expressed in *E. coli*. This resulted in 33 enzymes (55% of selected enzymes) that expressed well in *E. coli*, while 27 enzymes (45% of selected enzymes) failed in cloning and/or expression.

We then combined these 33 enzymes with 193 sequences from our in-house panel and ran a second round of

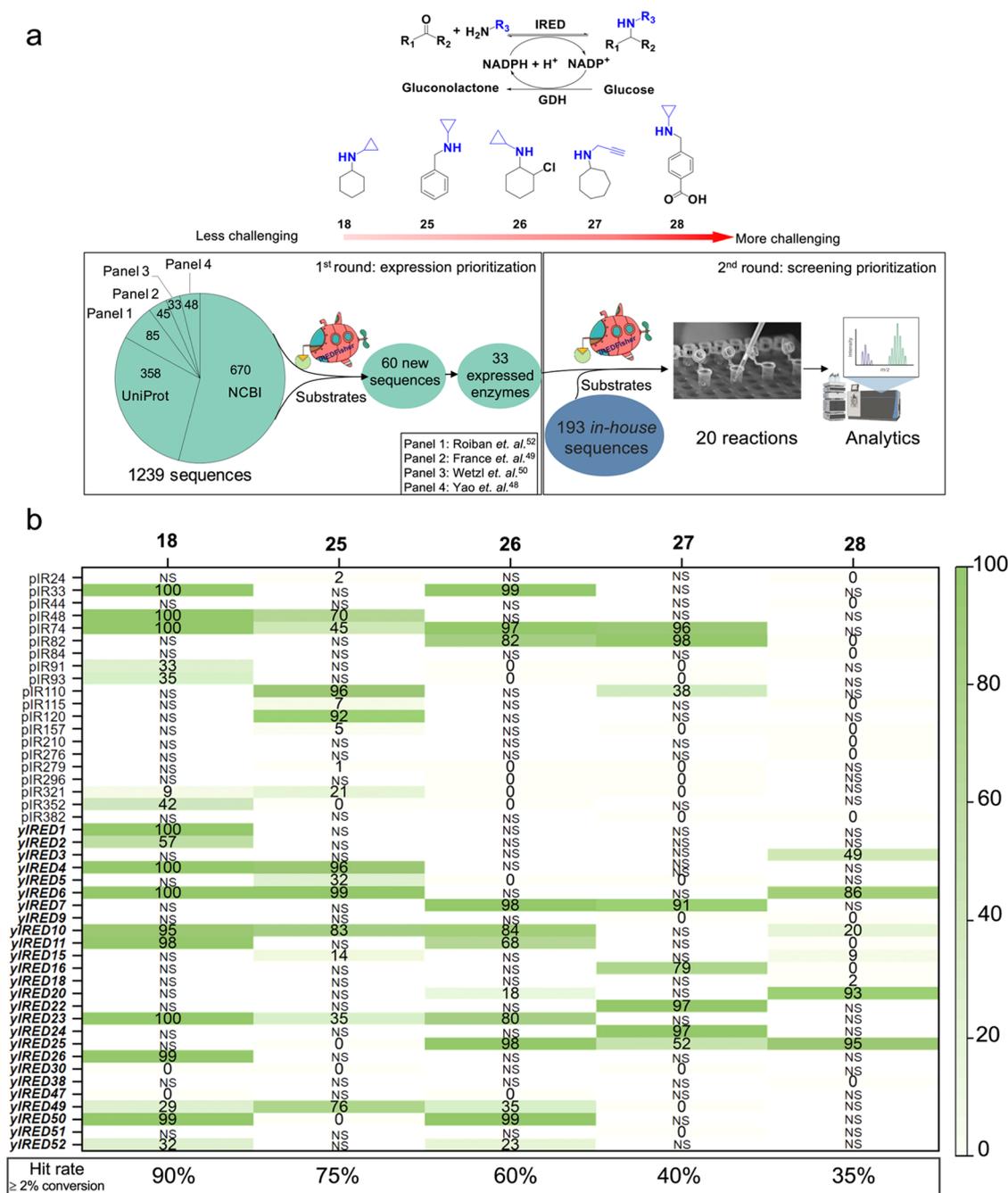


Figure 5. Validation of IREDFisher predictions in identifying hits for five chosen reactions. (a) IREDFisher was run in two rounds: in the first round, the sequences from external sources, public databases, and established panels were ranked by IREDFisher to prioritize those to be used for gene synthesis and enzyme expression. In the second round, 33 new enzymes with good expression in *Escherichia coli* and 193 in-house sequences were ranked by IREDFisher. The top 20 enzymes for each reaction were then investigated in the laboratory. The percentages of new enzymes and in-house enzymes are 65 and 35% for 18, 50 and 50% for 25, 45 and 55% for 26, 50 and 50% for 27, and 55 and 45% for 28. Compounds are shown in product forms. Amine nucleophile components are colored in blue. (b) The conversion of IREDFisher selected sequences toward products 18 and 25–28. Hit rates for each reaction were calculated using a conversion cutoff of 2%: Hit rate = $\frac{\text{number of enzymes } (\geq 2\% \text{ conversion})}{20} \times 100\%$. NS indicates that this sequence is not in the top 20 selected by IREDFisher and consequently was not tested in vitro. Sequences labeled pIR are from the in-house panel. Sequences labeled yIRED are new uncharacterized sequences and highlighted in bold and italic font.

IREDFisher ranking to prioritize these sequences in activity screening experiments. The top 20 IREDFishers obtained by IREDFisher for each targeted reaction were screened in vitro, and corresponding conversions to products were determined by GC or HPLC (Figure 5b). Reaction optimization was not carried out, as only the relative efficiency

of IREDFisher under standardized conditions was being evaluated. For target reactions 18 and 25–28, the hit rate decreases as the difficulty of the IRED reactions increases, with 2–11 hits (over 90% conversion) of corresponding substrates obtained. In a previous study, the hit rate of 26 (26b in a previous paper⁴⁸) and 27 (23d in a previous paper⁴⁸) by

a

b **Reminder:**
This page is refreshed every 30 seconds.
please bookmark this page and check back later.

UUID: 3e9c65f3-f81d-4fe4-95d6-37ef8c726cdf

Job name: test

Status: cross docking going on

notes: Fishing IREDs from Screened database, panel_optimization = False

Files:

aln_aln_complex.pdb	aln_complex_1a_imine_best_mode.pdb	1a_imine_summary_score.txt
aln_aln_IR035.pdb	aln_IR035_1a_imine_best_mode.pdb	1a_imine_ranking.txt
aln_aln_IR036.pdb	aln_IR036_1a_imine_best_mode.pdb	1a_imine_sort_rescore.csv
aln_aln_IR037.pdb	aln_IR037_1a_imine_best_mode.pdb	1a_imine_optimized_panel_sequence.fasta
		3e9c65f3-f81d-4fe4-95d6-37ef8c726cdf.zip

IRED 3D structures **IRED-substrate complex models** **Score files**

IRED	Refined score
IR046	-9.7
IR047	-9.6
IR083	-8.7
IR070	-8.3
IR081	-8.1
IR045	-7.9
IR048	-7.1
IR051	-7.1
IR037	-7.0
IR058	-6.9
IR057	-6.8
IR036	-6.3
IR040	-6.2
IR038	-6.1
IR050	-6.0
IR059	-6.0
IR080	-6.0
IR043	-5.9
IR063	-5.8
IR066	-5.8

Figure 6. Web interface of IREDFisher. (a) Job submission interface and input file examples. Left square: FASTA-formatted sequence file. Right square: generation of substrate structure and N atom index by Marvin JS (<https://marvinjs-demo.chemaxon.com/latest/>). (b) Output files. The file for sequence ranking (1a_imine_sort_rescore.csv in the example) is shown.

screening a panel of 384 IREDs was low (2.6% for **26** and 5.5% for **27**), whereas the hit rate in this study employing IREDFisher increased substantially to 60 and 40%, respectively. Notably, two hits, with conversion $\geq 90\%$, were found for the challenging product **28**, demonstrating that IREDFisher can find enzymes that catalyze reactions of pharmaceutical importance. Furthermore, by exploiting the resources from public databases, highly active IREDs that have not been reported previously were identified and multiple enzymes (e.g., yIRE10, yIRE23, yIRE25) were found to have broad substrate scope. The benefit of IREDFisher in prioritizing screening is clear, in this case reducing the experimental screening required from ≥ 1000 samples without IREDFisher virtual screening to only 20 samples.

IREDFisher Web Interface and Speed Test. Encouraged by the validation results using established screening data and our own experiments, our intention was to develop a user-friendly workflow that can be widely used by the biocatalysis

community. To avoid any requirement for the use of personal computational/programming skills, we have built an IREDFisher web server (<https://enzymeevolver.com/IREDFisher>) (Figure 6). We emphasize that the current methodology only applies to IREDs. For the input structure, the imine structure formed by the condensation of aldehyde or ketone and amine is required. This can be obtained by the free online structure drawing tool Marvin JS (<https://marvinjs-demo.chemaxon.com/latest/>) from ChemAxon (Figure 6a, example in the square on the right). The index of the N atom in the C=N bond is also required as an input, which can be generated using the Marvin interface. A FASTA-formatted sequence file is required if users intend to rank their own sequences (Figure 6a, example in the square on the left). After the job is submitted, output files such as three-dimensional structures of IREDs, substrate-bound complex structures, and scoring files are generated and added to the web page. By checking the sorted rescore file, the user can obtain an optimized small

IREDFisher screening panel with 20 recommended sequences by IREDFisher (Figure 6b). In addition, IREDFisher also allows users to rank the sequences from established panels and public databases (see the Methods section for details). Therefore, IREDFisher can be employed in ranking: (i) user-defined sequences; (ii) established IRED panels, e.g., Robian et al.,⁵⁶ France et al.,⁵² Wetzl et al.,⁵⁴ etc.; and (iii) IRED sequences collected from the public database in this study. The speed test showed that IREDFisher can perform the calculation of a user-defined panel with 85 sequences in 90 min. Moreover, it only takes ca 30 min to complete the ranking of the established IRED panels and sequences from public databases by using the structures already deposited in the IREDFisher web server. Consequently, based on simple data input, nonspecialists in bioinformatics can have access to workflow outputs using only a few mouse clicks without human interference, which makes it applicable for small IRED panel design.

DISCUSSION

The computational four-step workflow IREDFisher that aims to streamline the screening of enzyme panels was produced by integrating sequence analysis, structure modeling, and molecular docking. Although large numbers of protein structures have been released recently by AlphaFold,⁶² the structures obtained are not always applicable for the use of molecular docking. For many enzymes, the substrate-binding site is formed by monomers/dimers, but only monomeric structures were deposited in the database. For example, most of the new sequences used in this study were only modeled as monomer in the AlphaFold database. We have shown that the active site is similar between the dimer models generated by homology modeling and AlphaFold (Figure S2), while the former approach saves considerable computation time (3 min by homology modeling versus 30 min by AlphaFold per structure). This is also suggested by a previous publication.⁶³ Therefore, the template-based homology modeling is beneficial for the rapid building of structures in the IREDFisher workflow, allowing users to rank large numbers of sequences.

The workflow was applied by using a model enzyme IRED and subsequently validated using previously published screening data. IREDFisher demonstrated the power of retrieving the best hit found simply by screening a small panel of 20 sequences using the IREDFisher workflow. Experimental validation in which IREDFisher selects 20 enzymes from a pool of ~1400 sequences for 5 reactions, and with different levels of product complexity, further establishes the effectiveness of IREDFisher. IREDFisher ranks enzymes based on enzyme–substrate-binding interactions and does not consider features such as the charge distribution on the protein surface, which might, for example, be important for successful protein expression in *E. coli*. As such, poorly expressed proteins could rank better than proteins that are expressed well in *E. coli*. The workflow was run initially to identify sequences for gene synthesis and protein expression analysis. We then performed a second screening experiment to rank new IRED biocatalysts only with those sequences that expressed well in *E. coli*. The prediction of enantioselectivity by IREDFisher is not investigated in this study because this is challenging. That said, IREDFisher narrows the search to a small panel of IREDS such that screening those biocatalysts for enantioselectivity is no longer an onerous task.

It is noteworthy that most biocatalytic processes for drug production need to go through extensive reaction optimization

and mutagenesis approaches such as directed evolution to obtain a biocatalyst that fulfills the desired process requirements. Directed evolution relies on the screening of thousands of enzyme variants and usually involves several rounds of library generation and screening, resulting in large numbers of variants that require high-throughput screening (HTS) methods for their evaluation. As such, screening remains the major bottleneck in biocatalysis applications, and this places major burdens on the resources and time required to generate a suitable biocatalyst. By taking advantage of the IREDFisher computational workflow, we have shown that it is possible to identify suitable biocatalysts for simple and challenging substrates. If needed, these could be evolved further to fulfill process needs with reduced screening efforts. IREDFisher is therefore valuable, either as a standalone computational workflow to aid the rapid identification of biocatalysts or as a “front-end” workflow to accelerate directed evolution programs by selecting the prototype, catalytically competent enzyme scaffolds for a target reaction. However, we clarify that IREDFisher takes wild-type sequences as input, which limits its ability to “design” enzymes with non-natural functions to catalyze extremely challenging reactions.

IREDFisher is designed as a “fast prescreen method” to rank input sequences. As such, it has limitations. Protein dynamics and energy barrier calculations are not incorporated because this would be too time consuming for a fast prescreening tool for use by experimentalists. The IREDFisher scoring function works as a tool to rank input sequences instead of accurately predicting the conversion/activity of enzymes. We trained the scoring instead to deliver a maximum hit rate of the top 20 sequences because 20 screening experiments can be carried out in one batch with overnight incubation for rapid enzyme screening. That said, the IREDFisher workflow has the power to find highly active hits in the 20 recommended sequences. This indicates that a trade-off between speed and accuracy has been achieved to some extent. Currently, it is applicable to IREDS as the scoring function is designed specifically for this enzyme family. We have also tested IREDFisher on haloalkane dehydrogenases and compared our results with a previous study.^{39,40} Only one enzyme identified by this previous study was ranked in the top 20 sequences from IREDFisher. Clearly, generic substrate inputs and associated scoring functions will need to be explored to expand the utility of our workflow to different enzyme reactions. Specifically, reactive atoms in the substrate and catalytic residues in the enzyme should be provided by the user to help select the best binding pose after docking. Also, deep learning methods are promising approaches to train a model, e.g., by taking features from predicted binding modes such as distances and angles between key atoms, atom types, and charge distributions in the active site. This challenge, together with integration into Retro-Biocat,⁸ will be implemented to enable the process from the biocatalytic pathway design all the way through to the generation of small screening panels to allow for direct hit identification.

In summary, the development and use of IREDFisher have reduced the size of screening panels to only 20 enzymes, with significant savings in resource and time required to perform the experimental screening. For challenging substrates, where conversions with the wild-type enzyme may be low, the hits obtained could then be used as a parent sequence for subsequent directed evolution approaches. The workflow provides a user-friendly free web interface (<https://>

enzymeevolver.com/IREDFisher), requiring no experience in computational chemistry or programming, which users can implement with only a few “mouse clicks”. The expansion of the web tool to other enzyme classes is envisaged with more accurate scoring functions, taking advantage of machine learning and other approaches. This will assist in the identification of enzymes for transformations of interest from a plethora of continuously growing sequence data and, therefore, contribute to the increasing application of biocatalytic methods in the synthesis.

METHODS

IREDFisher Four-Step Panel Optimization Workflow.

Sequence Preprocessing. A FASTA-formatted sequence file given by the user was loaded into the workflow. Each individual sequence was generated by splitting the whole sequences. For sequence, we searched homologous proteins in the Protein Data Bank using MODELLER.⁵⁹ The one with the highest sequence identity was considered as the template protein of the query sequence. For imine reductases, only sequences with characterized IREDs (PDB codes: 3ZGY,⁶⁴ 4D3D,⁶⁵ 4D3S,⁶⁵ 4OQY,⁶⁶ 4OQZ,⁶⁶ 5A9T,⁶⁷ 5OCM,⁶⁸ 5OJL,⁶⁹ 6EOD,⁴⁴ 6JIT, 6JIZ, 6GRL, and 5G6R) as the template proteins were retained for the next step.

Structure Modeling. IREDs form homodimers with the active site located at the interface between the monomers, thus making it more challenging to build a model in an automated way.⁷⁰ The general sequence alignment between the template protein and query sequence was calculated only for one chain. To overcome this issue, the single-chain sequence alignment between each query sequence and the template protein was repeated once more by an in-house script to generate a double-chain sequence alignment file. To automate the modeling of the IRED dimeric form, a template protein needs to be specified. The imine reductase from *Streptosporangium roseum* (PDB code: 5OCM⁶⁸) is selected as the dimer template protein given its high resolution and the clear substrate and cofactor NADPH binding site. Then, models are generated and optimized using MODELLER⁵⁹ based on the modified double-chain sequence alignment. All models are evaluated based on the Ramachandran plot, and only those that have over 90% of the residues in the favorable/core region are defined as qualified models and subsequently are used for substrate docking.

Molecular Docking. To locate the active site easily, all three-dimensional structures obtained from the last step are aligned to the template structure 5OCM by the PyMol python library. The cofactor NADPH is directly placed into the cofactor binding site of each modeled structure using the reference NADPH from the crystal structure prior to substrate docking. The center of mass of the bound ligand 9RH in 5OCM is used as the center of the docking box. The docking box size is calculated by using script eBoxSize.pl,⁷¹ considering the gyration radius of the substrate, which improves the docking accuracy.⁷¹ Then, all aligned structures with the bound NADPH are prepared using AutoDockTools.⁷² To input the structure of the imine form of the substrate formed by the condensation of amine and ketone, the three-dimensional coordinates are generated by Babel⁷³ and then prepared using AutoDockTools.⁷² Multiple poses are generated by Autodock vina⁶¹ with exhaustiveness value 10.

Scoring and ranking. When the IRED reaction occurs, NADPH transfers a hydride from atom C4 to the imine

substrate.^{44,70,74} To obtain the most reliable pose, the distance between the C4 atom in NADPH and N in imine, the substrate is used as a criterion (Figure 2). Docking poses with the C4...N distance shorter than 3.5 Å are removed to avoid clashes. Poses with the C4...N distance longer than 6.0 Å are also removed because it is beyond the distance range for hydride transfer. The remaining poses having the closest C4...N distance are selected as the best model. To improve the scoring function, we take the catalytic mechanism, proton transfer, and hydride transfer⁷⁵ into consideration. It has been reported that acidic residues, basic residues, and a conserved histidine residue in the active site play important roles in the stabilization of the key intermediate in the IRED-catalyzed reaction.⁷⁵ Accordingly, we included three terms in the scoring function: the number of acidic residues N_{acidic} , the number of basic residues N_{basic} , and the presence of histidine residues to represent the effect enzymes have on reactions. The final scoring function after the optimization of hyperparameters is refined score = $4.0 \times \text{Vina score} + 1.0 \times N_{\text{acidic}} - 9.0 \times N_{\text{His}} + 9.0 \times N_{\text{basic}}$. In this equation, $N_{\text{acidic}}/N_{\text{basic}}$ is the total number of acidic/basic residues in the active site and N_{His} is the presence ($N_{\text{His}} = 1$) or absence of histidine ($N_{\text{His}} = 0$) residue in the active site. The weights for each component of the scoring function were tuned by running with a range of -10 to 10 using compounds 18–24 screened by Montgomery et al. IRED panel,⁵³ and the best weight was decided by the highest percentage of hits with conversion over 50% in the top 20 list. The scoring function was tested by Roiban et al. panel⁵⁶ for substrates 1–9 and France et al. panel⁵² for substrates 10–17. Only residues within 8 angstroms of the ligand are taken into account when we count the number of key residues.

IREDFisher Structure Databases. To help users design an IRED screening panel from scratch by the IREDFisher workflow, we collected published IRED panels: GSK panel in 2017,⁵⁶ MIB panel in 2018,⁵² MIB panel in 2020,⁵³ MIB metagenomic panel in 2021,⁴⁸ Roche panel in 2016,⁵⁵ and Peiyuan Yao panel⁵¹ in 2018. After sequence preprocessing step, a total of 451 published sequences were collected including the characterized IRED sequences.

To make use of the public database, we conducted homology search in UniProt⁶ and NCBI reference protein⁷⁶ databases using the characterized IREDs as seed sequences. Homologous sequences with sequence identity over 30%, coverage over 80%, and *E*-value under 10 compared to the corresponding seed sequence were downloaded. 1028 sequences were collected after removing duplicates, followed by a clustering with threshold 0.7 by cd-hit.⁷⁷ After sequence preprocessing, a total of 591 sequences were collected from the public database.

All of the above sequences were modeled and prepared for docking by the IREDFisher workflow. With the well-prepared structure database, IREDFisher runs much faster.

EXPERIMENTAL DETAILS FOR IN VITRO SCREENING

General (Chemicals and Enzymes). All chemicals and solvents were purchased from commercial sources such as Acros Organics (Geel, Belgium), Fluorochem (Hadfield, Derbyshire, UK), Alfa Aesar (Karlsruhe, Germany), and Sigma-Aldrich (Poole, Dorset, UK). Chemicals were used without additional purification. GC gases were obtained from BOC gases (Guildford, UK), and HPLC solvents were obtained from Honeywell (Seelze, Germany).

GC-MS analysis was performed using an Agilent 7890B Series GC with a 5977B MS-EI detector in positive mode at a constant He flow. HPLC analysis was performed using an Agilent 1200 Series with a UV detector Agilent 1200 Series. LC-MS analysis was performed using an Agilent 1260 Infinity II with an Agilent 6130 Quadrupole detector in positive mode using ESI as the ionization source.

Conversions were determined by achiral HPLC and GC-MS and calculated with respect to the carbonyl compound based on the convention used for reductive amination and by LC-MS with the calibration curve. Products were identified by LC-MS and by comparing retention time with standards.

■ ENZYME PRODUCTION AND EXPRESSION

All IREDs used in this study were obtained and expressed using the same conditions as outlined in <https://doi.org/10.1038/s41557-020-00606-w> Supporting Section S3.

■ SCREENING

Enzyme Screening for Target Product 18. Analytical scale reductive amination biotransformations were carried on 500 μL scale, adjusted to pH 8.0, containing 4 mg/mL of lyophilized powder of the supernatant of lysate IRED, 0.5 mg/mL of GDH (Codexis CDX-901), 0.5 mM of NADP⁺ (Prozomix), 20 mM of D-glucose, 2.5% of DMSO, 5 mM of cyclohexanone, and 10 equivalents of cyclopropylamine (1 M cyclopropylamine stock pH 8.0 adjusted), with the reaction volume made up to 500 μL in 100 mM tris buffer. The reaction mixture was incubated at 37 °C with shaking at 900 rpm (Eppendorf ThermoMixer) for 20 h. After 24 h, 500 μL was recovered from the reaction mixture and quenched with 20 μL of 10 M sodium hydroxide and extracted with 500 μL of dichloromethane (DCM). The organic phase was dried over anhydrous magnesium sulfate (MgSO_4) and analyzed by GC-MS.

■ ENZYME SCREENING FOR TARGET PRODUCT 25

Analytical scale reductive amination biotransformations were carried on a 500 μL scale adjusted to pH 8.0 containing 4 mg/mL of lyophilized powder of the supernatant of lysate IRED, 0.5 mg/mL of GDH (Codexis CDX-901), 0.5 mM of NADP⁺ (Prozomix), 40 mM of D-glucose, 5% of DMSO, 10 mM of benzaldehyde, and 5% and 10 equivalents of cyclopropylamine (1 M cyclopropylamine stock pH 8.0 adjusted), with the reaction volume made up to 500 μL in 100 mM tris buffer. The reaction mixture was incubated at 37 °C with shaking at 900 rpm (Eppendorf ThermoMixer) for 24 h. Following 24 h, the reaction was quenched by the addition of 1 volume of methanol. The mixture was centrifuged for 5 min at 13300 rpm, and the supernatant was recovered to be analyzed by HPLC and LC/MS.

■ ENZYME SCREENING FOR TARGET PRODUCT 26

Analytical scale reductive amination biotransformations were carried out on a 500 μL scale adjusted to pH 8.0 containing 4 mg/mL of lyophilized powder of the supernatant of lysate IRED, 0.5 mg/mL of GDH (Codexis CDX-901), 0.5 mM of NADP⁺ (Prozomix), 20 mM of D-glucose, 5 mM of 2-chlorocyclohexanone, and 10 equivalents of cyclopropylamine (1 M cyclopropylamine stock pH 8.0 adjusted), with the reaction volume made up to 500 μL in 100 mM tris buffer. The reaction mixture was incubated at 37 °C with shaking at 900

rpm (Eppendorf ThermoMixer) for 24 h. After 24 h, 250 μL was recovered from the reaction mixture and quenched with 20 μL of 10 M sodium hydroxide and extracted with 250 μL of dichloromethane (DCM) to extract the substrate and possible byproducts. The organic phase was dried over anhydrous magnesium sulfate (MgSO_4). The other 250 μL was extracted with 250 μL of dichloromethane (DCM) to extract the product. The two organic phases were combined, dried over anhydrous magnesium sulfate (MgSO_4), and analyzed by GC-MS.

■ ENZYME SCREENING FOR TARGET PRODUCT 27

Analytical scale reductive amination biotransformations were carried out on a 500 μL scale adjusted to pH 8.0 containing 4 mg/mL of lyophilized powder of the supernatant of lysate IRED, 0.5 mg/mL of GDH (Codexis CDX-901), 0.5 mM of NADP⁺ (Prozomix), 40 mM of D-glucose, 5% of DMSO, 10 mM of cycloheptanone, and 10 equivalents of propargylamine (1 M propargylamine stock pH 8.0 adjusted), with the reaction volume made up to 500 μL in 100 mM tris buffer. The reaction mixture was incubated at 37 °C with shaking at 900 rpm (Eppendorf ThermoMixer) for 24 h. After 24 h, 500 μL was recovered from the reaction mixture and quenched with 20 μL of 10 M sodium hydroxide and extracted with 500 μL of dichloromethane (DCM). The organic phase was dried over anhydrous magnesium sulfate (MgSO_4) and analyzed by GC-MS.

■ ENZYME SCREENING FOR TARGET PRODUCT 28

Previous test reactions were carried on at pH 8 as IREDs usually exhibit higher activities on slightly basic pH. However, reaction test 3 was done at pH 6 since the acid is partially deprotonated and the aim was to use pH values closer to formylbenzoic acid pKa to make it less negatively charged. Analytical scale reductive amination biotransformations were carried out on a 500 μL scale adjusted to pH 6.0 containing 4 mg/mL of lyophilized powder of the supernatant of lysate IRED, 0.5 mg/mL of GDH (Codexis CDX-901), 0.5 mM of NADP⁺ (Prozomix), 20 mM of D-glucose, 2.5% of DMSO, 5 mM of 4-formylbenzoic acid, and 20 equivalents of cyclopropylamine (1 M cyclopropylamine stock pH 6.0 adjusted), with the reaction volume made up to 500 μL in 100 mM 2-(N-morpholino) ethanesulfonic acid (MES) buffer. The reaction mixture was incubated at 37 °C with shaking at 200 rpm (incubator with orbital shaker) for 24 h and quenched by the addition of 1 volume of methanol. The mixture was centrifuged for 5 min at 13300 rpm, and the supernatant was recovered to be analyzed by HPLC and LC/MS.

■ ASSOCIATED CONTENT

Data Availability Statement

All data are within the article and the Supporting Information and are also available from the corresponding author upon reasonable request. IREDFisher is a free web server: <https://enzymeevolver.com/IREDFisher>. The source code was provided on GitHub (https://github.com/yuyuqi-design/iredfisher_src) and can be freely downloaded.

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acscatal.3c02278>.

Additional *in silico* screening details; experimental details; and materials and methods, including HPLC

and LC/MS chromatograms for the reactions conducted in this study and the sequences of all new putative imine reductases (PDF)

AUTHOR INFORMATION

Corresponding Authors

Yuqi Yu – Department of Chemistry, The University of Manchester, Manchester Institute of Biotechnology, Manchester M1 7DN, U.K.; Augmented Biologics Discovery & Design, Department of Biologics Engineering, BioPharmaceuticals R&D, AstraZeneca, Cambridge CB21 6GH, U.K.; Email: yuqi.yu@astrazeneca.com

Nicholas J. Turner – Department of Chemistry, The University of Manchester, Manchester Institute of Biotechnology, Manchester M1 7DN, U.K.; orcid.org/0000-0002-8708-0781; Email: nicholas.turner@manchester.ac.uk

Nigel S. Scrutton – Department of Chemistry, The University of Manchester, Manchester Institute of Biotechnology, Manchester M1 7DN, U.K.; orcid.org/0000-0002-4182-3500; Email: nigel.scrutton@manchester.ac.uk

Authors

Arnau Rué Casamajo – Department of Chemistry, The University of Manchester, Manchester Institute of Biotechnology, Manchester M1 7DN, U.K.; orcid.org/0000-0001-8932-8763

William Finnigan – Department of Chemistry, The University of Manchester, Manchester Institute of Biotechnology, Manchester M1 7DN, U.K.; orcid.org/0000-0002-1061-7800

Christian Schnepel – Department of Chemistry, The University of Manchester, Manchester Institute of Biotechnology, Manchester M1 7DN, U.K.; orcid.org/0000-0002-4331-2675

Rhys Barker – Department of Chemistry, The University of Manchester, Manchester Institute of Biotechnology, Manchester M1 7DN, U.K.

Charlotte Morrill – Department of Chemistry, The University of Manchester, Manchester Institute of Biotechnology, Manchester M1 7DN, U.K.

Rachel S. Heath – Department of Chemistry, The University of Manchester, Manchester Institute of Biotechnology, Manchester M1 7DN, U.K.

Leonardo De Maria – Medicinal Chemistry, Research and Early Development, Respiratory and Immunology (RI), BioPharmaceuticals R&D, AstraZeneca, Gothenburg 43150, Sweden; orcid.org/0000-0002-8061-4242

Complete contact information is available at: <https://pubs.acs.org/10.1021/acscatal.3c02278>

Author Contributions

N.S.S., N.J.T., and L.D.M. supervised the project. N.S.S., N.J.T., Y.Y., and L.D.M. devised the concept. Y.Y., W.F., and R.B. performed the construction of the workflow, web server, and database. Y.Y. performed the in silico validation of the workflow. A.R.C., Y.Y., C.S., and R.S.H. performed the experiment. C.M. performed the organic synthesis. Y.Y., A.R.C., N.S.S., N.J.J., and L.D.M. wrote the manuscript and generated the figures.

Notes

The authors declare the following competing financial interest(s): NSS is co-founder, executive director and share holder of C3 Biotechnologies Ltd which is active in engineering biology for synthetic fuels production.

ACKNOWLEDGMENTS

The authors thank Research IT and the use of the Computational Shared Facility at The University of Manchester. This study was funded by the EPSRC, BBSRC, and AstraZeneca plc for funding under the Prosperity Partnership EP/S005226/1.

REFERENCES

- (1) Clouthier, C. M.; Pelletier, J. N. Expanding the organic toolbox: a guide to integrating biocatalysis in synthesis. *Chem. Soc. Rev.* **2012**, *41*, 1585–1605.
- (2) Sheldon, R. A.; Woodley, J. M. Role of Biocatalysis in Sustainable Chemistry. *Chem. Rev.* **2018**, *118*, 801–838.
- (3) Panke, S.; Held, M.; Wubbolts, M. Trends and innovations in industrial biocatalysis for the production of fine chemicals. *Curr. Opin. Biotechnol.* **2004**, *15*, 272–279.
- (4) Kohls, H.; Steffen-Munsberg, F.; Hohne, M. Recent achievements in developing the biocatalytic toolbox for chiral amine synthesis. *Curr. Opin. Chem. Biol.* **2014**, *19*, 180–192.
- (5) Wu, S.; Snajdrova, R.; Moore, J. C.; Baldenius, K.; Bornscheuer, U. T. Biocatalysis: Enzymatic Synthesis for Industrial Applications. *Angew. Chem., Int. Ed.* **2021**, *60*, 88–119.
- (6) UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489.
- (7) Sayers, E. W.; Beck, J.; Bolton, E. E.; Bourexis, D.; Brister, J. R.; Canese, K.; Comeau, D. C.; Funk, K.; Kim, S.; Klimke, W.; Marchler-Bauer, A.; Landrum, M.; Lathrop, S.; Lu, Z.; Madden, T. L.; O’Leary, N.; Phan, L.; Rangwala, S. H.; Schneider, V. A.; Skripchenko, Y.; Wang, J.; Ye, J.; Trawick, B. W.; Pruitt, K. D.; Sherry, S. T. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2021**, *49*, D10–D17.
- (8) 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.* **2021**, *4*, 98–104.
- (9) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-Assisted Retrosynthesis Based on Molecular Similarity. *ACS Cent. Sci.* **2017**, *3*, 1237–1245.
- (10) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem., Int. Ed.* **2016**, *55*, 5904–5937.
- (11) Cui, Y.; Wang, Y.; Tian, W.; Bu, Y.; Li, T.; Cui, X.; Zhu, T.; Li, R.; Wu, B. Development of a versatile and efficient C–N lyase platform for asymmetric hydroamination via computational enzyme redesign. *Nat. Catal.* **2021**, *4*, 364–373.
- (12) Bell, S. G.; Chen, X.; Xu, F.; Rao, Z.; Wong, L. L. Engineering substrate recognition in catalysis by cytochrome P450cam. *Biochem. Soc. Trans.* **2003**, *31*, 558–562.
- (13) Dunn, B. J.; Khosla, C. Engineering the acyltransferase substrate specificity of assembly line polyketide synthases. *J. R. Soc., Interface* **2013**, *10*, No. 20130297.
- (14) Ali, M.; Ishqi, H. M.; Husain, Q. Enzyme engineering: Reshaping the biocatalytic functions. *Biotechnol. Bioeng.* **2020**, *117*, 1877–1894.
- (15) Quijano-Rubio, A.; Yeh, H. W.; Park, J.; Lee, H.; Langan, R. A.; Boyken, S. E.; Lajoie, M. J.; Cao, L.; Chow, C. M.; Miranda, M. C.; Wi, J.; Hong, H. J.; Stewart, L.; Oh, B.-H.; Baker, D. De novo design of modular and tunable protein biosensors. *Nature* **2021**, *591*, 482–487.
- (16) Vorobieva, A. A.; White, P.; Liang, B.; Horne, J. E.; Bera, A. K.; Chow, C. M.; Gerben, S.; Marx, S.; Kang, A.; Stiving, A. Q.; Harvey, S. R.; Marx, D. C.; Khan, G. N.; Fleming, K. G.; Wysocki, V. H.;

- Brockwell, D. J.; Tamm, L. K.; Radford, S. E.; Baker, D. De novo design of transmembrane beta barrels. *Science* **2021**, *371*, No. eabc8182.
- (17) Klima, J. C.; Doyle, L. A.; Lee, J. D.; Rappleye, M.; Gagnon, L. A.; Lee, M. Y.; Barros, E. P.; Vorobieva, A. A.; Dou, J.; Bremner, S.; Quon, J. S.; Chow, C. M.; Carter, L.; Mack, D. L.; Amaro, R. E.; Vaughan, J. C.; Berndt, A.; Stoddard, B. L.; Baker, D. Incorporation of sensing modalities into de novo designed fluorescence-activating proteins. *Nat. Commun.* **2021**, *12*, 856.
- (18) Sahtoe, D. D.; Coscia, A.; Mustafaoglu, N.; Miller, L. M.; Olal, D.; Vulovic, I.; Yu, T. Y.; Goreschnik, I.; Lin, Y. R.; Clark, L.; Busch, F.; Stewart, L.; Wysocki, V. H.; Ingber, D. E.; Abraham, J.; Baker, D. Transferrin receptor targeting by de novo sheet extension. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, 1–8.
- (19) Stourac, J.; Vavra, O.; Kokkonen, P.; Filipovic, J.; Pinto, G.; Brezovsky, J.; Damborsky, J.; Bednar, D. Caver Web 1.0: identification of tunnels and channels in proteins and analysis of ligand transport. *Nucleic Acids Res.* **2019**, *47*, W414–W422.
- (20) Sumbalova, L.; Stourac, J.; Martinek, T.; Bednar, D.; Damborsky, J. HotSpot Wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information. *Nucleic Acids Res.* **2018**, *46*, W356–W362.
- (21) Angles, R.; Arenas-Salinas, M.; Garcia, R.; Reyes-Suarez, J. A.; Pohl, E. GSP4PDB: a web tool to visualize, search and explore protein-ligand structural patterns. *BMC Bioinf.* **2020**, *21*, 85.
- (22) Viet Hung, L.; Caprari, S.; Bizai, M.; Toti, D.; Polticelli, F. LIBRA: Llgand Binding site Recognition Application. *Bioinformatics* **2015**, *31*, 4020–4022.
- (23) Toti, D.; Viet Hung, L.; Tortosa, V.; Brandi, V.; Polticelli, F. LIBRA-WA: a web application for ligand binding site detection and protein function recognition. *Bioinformatics* **2018**, *34*, 878–880.
- (24) Lilley, D. M.; Clegg, R. M.; Diekmann, S.; Seeman, N. C.; von Kitzing, E.; Hagerman, P. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). A nomenclature of junctions and branchpoints in nucleic acids. Recommendations 1994. *Eur. J. Biochem.* **1995**, *230*, 1–2.
- (25) Dalkiran, A.; Rifaioğlu, A. S.; Martin, M. J.; Cetin-Atalay, R.; Atalay, V.; Dogan, T. ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinf.* **2018**, *19*, 334.
- (26) Shen, H. B.; Chou, K. C. EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. *Biochem. Biophys. Res. Commun.* **2007**, *364*, 53–59.
- (27) Li, Y.; Wang, S.; Umarov, R.; Xie, B.; Fan, M.; Li, L.; Gao, X. DEEPred: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics* **2018**, *34*, 760–769.
- (28) Memon, S. A.; Khan, K. A.; Naveed, H. HECNet: a hierarchical approach to enzyme function classification using a Siamese Triplet Network. *Bioinformatics* **2020**, *36*, 4583–4589.
- (29) Zhang, T.; Tian, Y.; Yuan, L.; Chen, F.; Ren, A.; Hu, Q. N. Bio2Rxn: sequence-based enzymatic reaction predictions by a consensus strategy. *Bioinformatics* **2020**, *36*, 3600–3601.
- (30) Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G. A.; Sonnhammer, E. L. L.; Tosatto, S. C. E.; Paladin, L.; Raj, S.; Richardson, L. J.; Finn, R. D.; Bateman, A. Pfam: The protein families database in 2021. *Nucleic Acids Res.* **2021**, *49*, D412–D419.
- (31) Blum, M.; Chang, H. Y.; Chuguransky, S.; Grego, T.; Kandasamy, S.; Mitchell, A.; Nuka, G.; Paysan-Lafosse, T.; Qureshi, M.; Raj, S.; Richardson, L.; Salazar, G. A.; Williams, L.; Bork, P.; Bridge, A.; Gough, J.; Haft, D. H.; Letunic, I.; Marchler-Bauer, A.; Mi, H.; Natale, D.; Necci, M.; Orengo, C. A.; Pandurangen, A. P.; Roivore, C.; Sigrist, C. J. A.; Sillitoe, I.; Thanki, N.; Thomas, P. D.; Tosatto, S. C. E.; Wu, C. H.; Bateman, A.; Finn, R. D. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **2021**, *49*, D344–D354.
- (32) Hon, J.; Borko, S.; Stourac, J.; Prokop, Z.; Zendulka, J.; Bednar, D.; Martinek, T.; Damborsky, J. EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities. *Nucleic Acids Res.* **2020**, *48*, W104–W109.
- (33) Wu, C. H.; Nikolskaya, A.; Huang, H.; Yeh, L. S.; Natale, D. A.; Vinayaka, C. R.; Hu, Z. Z.; Mazumder, R.; Kumar, S.; Kourtesis, P.; Ledley, R. S.; Suzek, B. E.; Arminski, L.; Chen, Y.; Zhang, J.; Cardenas, J. L.; Chung, S.; Castro-Alvarez, J.; Dinkov, G.; Barker, W. C. PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res.* **2004**, *32*, D112–D114.
- (34) Lu, L.; Qian, Z.; Cai, Y. D.; Li, Y. ECS: an automatic enzyme classifier based on functional domain composition. *Comput. Biol. Chem.* **2007**, *31*, 226–232.
- (35) Alborzi, S. Z.; Devignes, M. D.; Ritchie, D. W. ECDomain-Miner: discovering hidden associations between enzyme commission numbers and Pfam domains. *BMC Bioinf.* **2017**, *18*, 107.
- (36) Sigrist, C. J. A.; Cerutti, L.; de Castro, E.; Langendijk-Genevaux, P. S.; Bulliard, V.; Bairoch, A.; Hulo, N. PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.* **2010**, *38*, D161–166.
- (37) de Rond, T.; Asay, J. E.; Moore, B. S. Co-occurrence of enzyme domains guides the discovery of an oxazolone synthetase. *Nat. Chem. Biol.* **2021**, *17*, 794–799.
- (38) Mak, W. S.; Tran, S.; Marcheschi, R.; Bertolani, S.; Thompson, J.; Baker, D.; Liao, J. C.; Siegel, J. B. Integrative genomic mining for enzyme function to enable engineering of a non-natural biosynthetic pathway. *Nat. Commun.* **2015**, *6*, No. 10005.
- (39) Vasina, M.; Vanacek, P.; Hon, J.; Kovar, D.; Faldynova, H.; Kunka, A.; Buryška, T.; Badenhorst, C. P. S.; Mazurenko, S.; Bednar, D.; Stavrakis, S.; Bornscheuer, U. T.; deMello, A.; Damborsky, J.; Prokop, Z. Advanced database mining of efficient haloalkane dehalogenases by sequence and structure bioinformatics and microfluidics. *Chem. Catal.* **2022**, *2*, 2704–2725.
- (40) Vanacek, P.; Sebestova, E.; Babkova, P.; Bidmanova, S.; Daniel, L.; Dvorak, P.; Stepankova, V.; Chaloupkova, R.; Brezovsky, J.; Prokop, Z.; Damborsky, J. Exploration of Enzyme Diversity by Integrating Bioinformatics with Expression Analysis and Biochemical Characterization. *ACS Catal.* **2018**, *8*, 2402–2412.
- (41) Lobb, B.; Doxey, A. C. Novel function discovery through sequence and structural data mining. *Curr. Opin. Struct. Biol.* **2016**, *38*, 53–61.
- (42) Ghislieri, D.; Turner, N. J. Biocatalytic Approaches to the Synthesis of Enantiomerically Pure Chiral Amines. *Top. Catal.* **2014**, *57*, 284–300.
- (43) Wu, Z.; Du, S.; Gao, G.; Yang, W.; Yang, X.; Huang, H.; Chang, M. Secondary amines as coupling partners in direct catalytic asymmetric reductive amination. *Chem. Sci.* **2019**, *10*, 4509–4514.
- (44) Sharma, M.; Mangas-Sanchez, J.; France, S. P.; Aleku, G. A.; Montgomery, S. L.; Ramsden, J. L.; Turner, N. J.; Grogan, G. A mechanism for reductive amination catalyzed by fungal reductive aminases. *ACS Catal.* **2018**, *8*, 11534–11541.
- (45) Devine, P. N.; Howard, R. M.; Kumar, R.; Thompson, M. P.; Truppo, M. D.; Turner, N. J. Extending the application of biocatalysis to meet the challenges of drug development. *Nat. Rev. Chem.* **2018**, *2*, 409–421.
- (46) Aleku, G. A.; France, S. P.; Man, H.; Mangas-Sanchez, J.; Montgomery, S. L.; Sharma, M.; Leipold, F.; Hussain, S.; Grogan, G.; Turner, N. J. A reductive aminase from *Aspergillus oryzae*. *Nat. Chem.* **2017**, *9*, 961–969.
- (47) Mitsukura, K.; Suzuki, M.; Tada, K.; Yoshida, T.; Nagasawa, T. Asymmetric synthesis of chiral cyclic amine from cyclic imine by bacterial whole-cell catalyst of enantioselective imine reductase. *Org. Biomol. Chem.* **2010**, *8*, 4533–4535.
- (48) Marshall, J. R.; Yao, P.; Montgomery, S. L.; Finnigan, J. D.; Thorpe, T. W.; Palmer, R. B.; Mangas-Sanchez, J.; Duncan, R. A. M.; Heath, R. S.; Graham, K. M.; Cook, D. J.; Charnock, S. J.; Turner, N. J. Screening and characterization of a diverse panel of metagenomic imine reductases for biocatalytic reductive amination. *Nat. Chem.* **2021**, *13*, 140–148.
- (49) Li, H.; Luan, Z.-J.; Zheng, G.-W.; Xu, J.-H. Efficient Synthesis of Chiral Indolines using an Imine Reductase from *Paenibacillus lactis*. *Adv. Synth. Catal.* **2015**, *357*, 1692–1696.

- (50) Matzel, P.; Gand, M.; Höhne, M. One-step asymmetric synthesis of (R)- and (S)-rasagiline by reductive amination applying imine reductases. *Green Chem.* **2017**, *19*, 385–389.
- (51) Yao, P.; Xu, Z.; Yu, S.; Wu, Q.; Zhu, D. Imine Reductase-Catalyzed Enantioselective Reduction of Bulky α,β -Unsaturated Imines en Route to a Pharmaceutically Important Morphinan Skeleton. *Adv. Synth. Catal.* **2019**, *361*, 556–561.
- (52) France, S. P.; Howard, R. M.; Steflik, J.; Weise, N. J.; Mangas-Sanchez, J.; Montgomery, S. L.; Crook, R.; Kumar, R.; Turner, N. J. Identification of Novel Bacterial Members of the Imine Reductase Enzyme Family that Perform Reductive Amination. *ChemCatChem* **2018**, *10*, 510–514.
- (53) Montgomery, S. L.; Pushpanath, A.; Heath, R. S.; Marshall, J. R.; Klemstein, U.; Galman, J. L.; Woodlock, D.; Bisagni, S.; Taylor, C. J.; Mangas-Sanchez, J.; Ramsden, J. I.; Dominguez, B.; Turner, N. J. Characterization of imine reductases in reductive amination for the exploration of structure-activity relationships. *Sci. Adv.* **2020**, *6*, No. eaay9320.
- (54) Wetzl, D.; Berrera, M.; Sandon, N.; Fishlock, D.; Ebeling, M.; Müller, M.; Hanlon, S.; Wirz, B.; Iding, H. Expanding the Imine Reductase Toolbox by Exploring the Bacterial Protein-Sequence Space. *ChemBioChem* **2015**, *16*, 1749–1756.
- (55) Wetzl, D.; Gand, M.; Ross, A.; Müller, H.; Matzel, P.; Hanlon, S. P.; Müller, M.; Wirz, B.; Höhne, M.; Iding, H. Asymmetric Reductive Amination of Ketones Catalyzed by Imine Reductases. *ChemCatChem* **2016**, *8*, 2023–2026.
- (56) Roiban, G.-D.; Kern, M.; Liu, Z.; Hyslop, J.; Tey, P. L.; Levine, M. S.; Jordan, L. S.; Brown, K. K.; Hadi, T.; Ihnken, L. A. F.; Brown, M. J. B. Efficient Biocatalytic Reductive Aminations by Extending the Imine Reductase Toolbox. *ChemCatChem* **2017**, *9*, 4475–4479.
- (57) Schober, M.; MacDermaid, C.; Ollis, A. A.; Chang, S.; Khan, D.; Hosford, J.; Latham, J.; Ihnken, L. A. F.; Brown, M. J. B.; Fuerst, D.; Sangane, M. J.; Roiban, G.-D. Chiral synthesis of LSD1 inhibitor GSK2879552 enabled by directed evolution of an imine reductase. *Nat. Catal.* **2019**, *2*, 909–915.
- (58) Kumar, R.; Karmilowicz, M. J.; Burke, D.; Burns, M. P.; Clark, L. A.; Connor, C. G.; Cordi, E.; Do, N. M.; Doyle, K. M.; Hoagland, S.; Lewis, C. A.; Mangan, D.; Martinez, C. A.; McInturff, E. L.; Meldrum, K.; Pearson, R.; Steflik, J.; Rane, A.; Weaver, J. Biocatalytic reductive amination from discovery to commercial manufacturing applied to abrocitinib JAK1 inhibitor. *Nat. Catal.* **2021**, *4*, 775–782.
- (59) Webb, B.; Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinf.* **2016**, *54*, 5–6.
- (60) Sánchez, R.; Sali, A. Comparative protein structure modeling as an optimization problem. *J. Mol. Struct.: THEOCHEM* **1997**, *398-399*, 489–496.
- (61) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461.
- (62) Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Zidek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; Velankar, S.; Kleywegt, G. J.; Bateman, A.; Evans, R.; Pritzel, A.; Figurnov, M.; Ronneberger, O.; Bates, R.; Kohl, S. A. A.; Potapenko, A.; Ballard, A. J.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Clancy, E.; Reiman, D.; Petersen, S.; Senior, A. W.; Kavukcuoglu, K.; Birney, E.; Pushmeet, K.; Jumper, J.; Hassabis, D. Highly accurate protein structure prediction for the human proteome. *Nature* **2021**, *596*, 590–596.
- (63) Lee, C.; Su, B. H.; Tseng, Y. J. Comparative studies of AlphaFold, RoseTTAFold and Modeller: a case study involving the use of G-protein-coupled receptors. *Briefings Bioinf.* **2022**, *23*, 109–120.
- (64) Rodríguez-Mata, M.; Frank, A.; Wells, E.; Leipold, F.; Turner, N. J.; Hart, S.; Turkenburg, J. P.; Grogan, G. Structure and activity of NADPH-dependent reductase Q1EQE0 from *Streptomyces kanamyceticus*, which catalyses the R-selective reduction of an imine substrate. *ChemBioChem* **2013**, *14*, 1372–1379.
- (65) Man, H.; Wells, E.; Hussain, S.; Leipold, F.; Hart, S.; Turkenburg, J. P.; Turner, N. J.; Grogan, G. Structure, Activity and Stereoselectivity of NADPH-Dependent Oxidoreductases Catalysing the S-Selective Reduction of the Imine Substrate 2-Methylpyrrolone. *ChemBioChem* **2015**, *16*, 1052–1059.
- (66) Huber, T.; Schneider, L.; Präg, A.; Gerhardt, S.; Einsle, O.; Müller, M. Direct Reductive Amination of Ketones: Structure and Activity of S-Selective Imine Reductases from *Streptomyces*. *ChemCatChem* **2014**, *6*, 2248–2252.
- (67) Aleku, G. A.; Man, H.; France, S. P.; Leipold, F.; Hussain, S.; Toca-Gonzalez, L.; Marchington, R.; Hart, S.; Turkenburg, J. P.; Grogan, G.; Turner, N. J. Stereoselectivity and Structural Characterization of an Imine Reductase (IRED) from *Amycolatopsis orientalis*. *ACS Catal.* **2016**, *6*, 3880–3889.
- (68) Lenz, M.; Fademrecht, S.; Sharma, M.; Pleiss, J.; Grogan, G.; Nestl, B. M. New imine-reducing enzymes from beta-hydroxyacid dehydrogenases by single amino acid substitutions. *Protein Eng., Des. Sel.* **2018**, *31*, 109–120.
- (69) France, S. P.; Aleku, G. A.; Sharma, M.; Mangas-Sanchez, J.; Howard, R. M.; Steflik, J.; Kumar, R.; Adams, R. W.; Slabu, I.; Crook, R.; Grogan, G.; Wallace, T. W.; Turner, N. J. Biocatalytic Routes to Enantiomerically Enriched Dibenz[*c,e*]azepines. *Angew. Chem., Int. Ed.* **2017**, *56*, 15589–15593.
- (70) Mangas-Sanchez, J.; France, S. P.; Montgomery, S. L.; Aleku, G. A.; Man, H.; Sharma, M.; Ramsden, J. I.; Grogan, G.; Turner, N. J. Imine reductases (IREDs). *Curr. Opin. Chem. Biol.* **2017**, *37*, 19–25.
- (71) Feinstein, W. P.; Brylinski, M. Calculating an optimal box size for ligand docking and virtual screening against experimental and predicted binding pockets. *J. Cheminf.* **2015**, *7*, No. 18.
- (72) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009**, *30*, 2785–2791.
- (73) O’Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.
- (74) Stockinger, P.; Roth, S.; Müller, M.; Pleiss, J. Systematic Evaluation of Imine-Reducing Enzymes: Common Principles in Imine Reductases, beta-Hydroxy Acid Dehydrogenases, and Short-Chain Dehydrogenases/ Reductases. *ChemBioChem* **2020**, *21*, 2689–2695.
- (75) Fademrecht, S.; Scheller, P. N.; Nestl, B. M.; Hauer, B.; Pleiss, J. Identification of imine reductase-specific sequence motifs. *Proteins* **2016**, *84*, 600–610.
- (76) Pruitt, K. D.; Tatusova, T.; Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **2007**, *35*, D61–65.
- (77) Li, W.; Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659.