Ligify^{DB}: An interactive platform for predicted small molecule biosensors

Simon d'Oelsnitz* [1,2,3], Nicole Zhao[1,3], Pranay Talla[1,3], Jio Jeong[4], Joshua D. Love[5], Michael Springer[1,2,3], Pamela A. Silver[1,2,3].

Affiliations

[1] Synthetic Biology HIVE, Department of Systems Biology, Harvard Medical School, Boston, MA, 02115, USA

[2] Wyss Institute for Biologically Inspired Engineering at Harvard University, Boston, MA, 02115, USA.

[3] Department of Systems Biology, Harvard Medical School, Boston, MA, 02115, USA.

[4] Department of Chemistry and Chemical Biology, Harvard University, Boston, MA, 02115, USA.

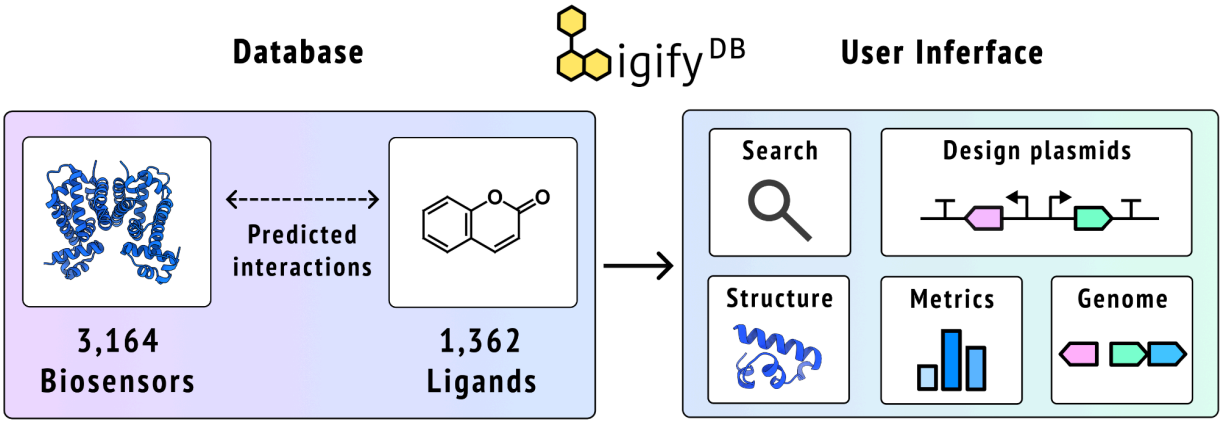
[5] Independent Web Developer. Bentonville, AR, 72712, USA

* To whom correspondence should be addressed: Simon d'Oelsnitz, simonsnitz@gmail.com

Abstract

Prokaryotic transcription factors (TFs) are indispensable small molecule biosensors with broad utility in biotechnology, but only a small fraction have been characterized. To address this issue, recent bioinformatic methods have leveraged information from enzyme reaction databases to predict effector molecules for a given TF. Here, we extend this approach by systematically evaluating 10,965 small molecules within the Rhea enzyme reaction database for association to TFs. Ultimately, this yields the Ligify^{DB} database, which contains predicted associations for 3,164 unique TFs to 1,362 unique small molecules. We then create an interactive web application (https://ligify.groov.bio) to access and query the data, which contains detailed information on chemical ligand structures, TF structures, metadata, genome context, and prediction confidence metrics. The open-access database and user interface for predicted TFs reported herein aims to facilitate the systematic characterization and utility of genetic biosensors for chemical-control of biological systems.

Graphic abstract



Introduction

Genetically-encoded small molecule biosensors are core to many applications in biotechnology, ranging from high-throughput screening, diagnostics, responsive materials, and live-cell metabolite tracking[1–4]. Among biosensor modalities, prokaryotic ligand-inducible transcription factors (TFs) are one of the most widely used due to their broad host compatibility, compact size, and high sensitivity and dynamic range[5]. TFs regulate gene expression by undergoing ligand-induced conformational changes that modulate their DNA-binding affinity, thereby either promoting RNA polymerase recruitment or blocking its access to the promoter.

Despite their acknowledged utility, identifying a TF with desired ligand specificity remains challenging. Large genome modeling efforts have driven the characterization of TFs in model organisms, producing databases such as RegulonDB, BioCyc, and CoryneRegNet[6–8]. However, ligands for the vast majority of TFs outside these few organisms are still largely unknown. To help close this gap, genome context has proven to be effective at guiding the prediction of effector ligands for a given TF[9]. For example, a TetR-family regulator co-transcribed with an enzyme that glycosylates the antibiotic kijanimicin was found to bind specifically to kijanimicin as a ligand[10].

Using genome context alone, recent software tools have automated the prediction of TF-ligand interactions. The Ligify and TFBMiner tools perform this prediction task by leveraging data from the enzyme reaction databases Rhea and KEGG, respectively, assuming that TFs encoded nearby an enzyme that acts on a given molecule might also respond to that same molecule[11–14]. These tools aim to expand the knowledgebase of TF-ligand interactions contained within databases, such as groov^{DB}, SensBio, RegPrecise, and the detectable molecules collection[15–18]. While useful, the tools suffer from several issues limiting their potential. All predictions must be computed on demand, which is slow, computationally expensive, limits use for bioinformaticians, and complicates a comprehensive analysis of prediction performance. Additionally, user interfaces are limited to the command line or simple text-based web applications, which delivers a poor user interface and hinders data interpretation.

Here, we report Ligify^{DB} (https://ligify.groov.bio), an interactive and extensible platform for predicted TF biosensors that overcomes accessibility, performance, and usability issues with existing TF prediction tools. Generated by running the Ligify workflow on over 10,000 chemicals in the Rhea database, Ligify^{DB} contains 3,164 unique TFs predicted to respond to 1,362 unique small molecules. Analyses indicate the database content covers a wide distribution of TFs from phylogenetically diverse hosts, and includes a range of metabolites such as sugars, metals, amino acids, and aromatics. We then create an interactive interface to easily query predictions via text search, chemical similarity search, and attribute filtering, and visualize TF entry properties, such as their 3D protein structure, 2D chemical ligand structures, operons, predicted promoters, and prediction quality metrics. Finally, a plasmid designer tool facilitates the creation of reporter plasmids for experimental validation of predictions.

Results

Creating the database

We sought to improve the accessibility and analysis of TF biosensor predictions by pre-computing all possible results using Ligify. As described in the original publication, the Ligify workflow starts by using a small molecule input to fetch associated enzyme reactions from the Rhea database[14] (**Figure 1**). Bacterial enzymes annotated to perform these reactions are fetched from Uniprot, and highly similar enzymes are filtered to reduce redundancy. Genes encoded within each enzyme's operon are retrieved from the NCBI Gene Database, and their annotations are scanned for the terms "regulator", "repressor", and "activator" to identify TFs. If a TF is found, the corresponding promoter element is extracted, under the general assumption that TFs tend to regulate their own expression. If TFs are identified, the output of the Ligify workflow includes the TF gene and promoter, genes within the operon, associated literature, a rank calculated based on a set of heuristics, metadata links, and all possible candidate ligands.

To create a database containing all possible predictions from Ligify, we started by fetching all unique small molecules from the Rhea database in ChEBI format, which was 13,972 at the time of database generation. Each molecule was then converted into a SMILES format compatible with Ligify, which returned 10,965 molecules since some ChEBI entries did not have an associated SMILES identifier. After filtering out generic molecules, such as "water" or "hydron", the Ligify workflow was performed on each molecule using default parameters (**see Methods**). Ultimately, 1,685 unique molecules were predicted to be associated with a total of 3,215 unique TFs. To reduce false positives we performed two filtering steps. First, we removed TFs smaller than 80 amino acids, since some of the shortest TFs known to bind DNA and a small molecule are around 90 amino acids long[19,20]. Next, we removed common co-factors, such as ATP and Coenzyme A, since these are overrepresented and unlikely to be the true ligands of any given TF (**see Methods**). Ultimately, this yielded 3,164 unique TFs associated with 1,362 unique molecules. To finalize the database, we (1) removed unnecessary data fields, such as operon sequences, to reduce size, (2) added the protein regulator sequence, (3) added Uniprot links to each regulator when possible, and (4) added links to homologous characterized TFs in the groov^{DB} database[16]. Ultimately, the final database is 12 megabytes and contains 13,435 hypothetical interactions.

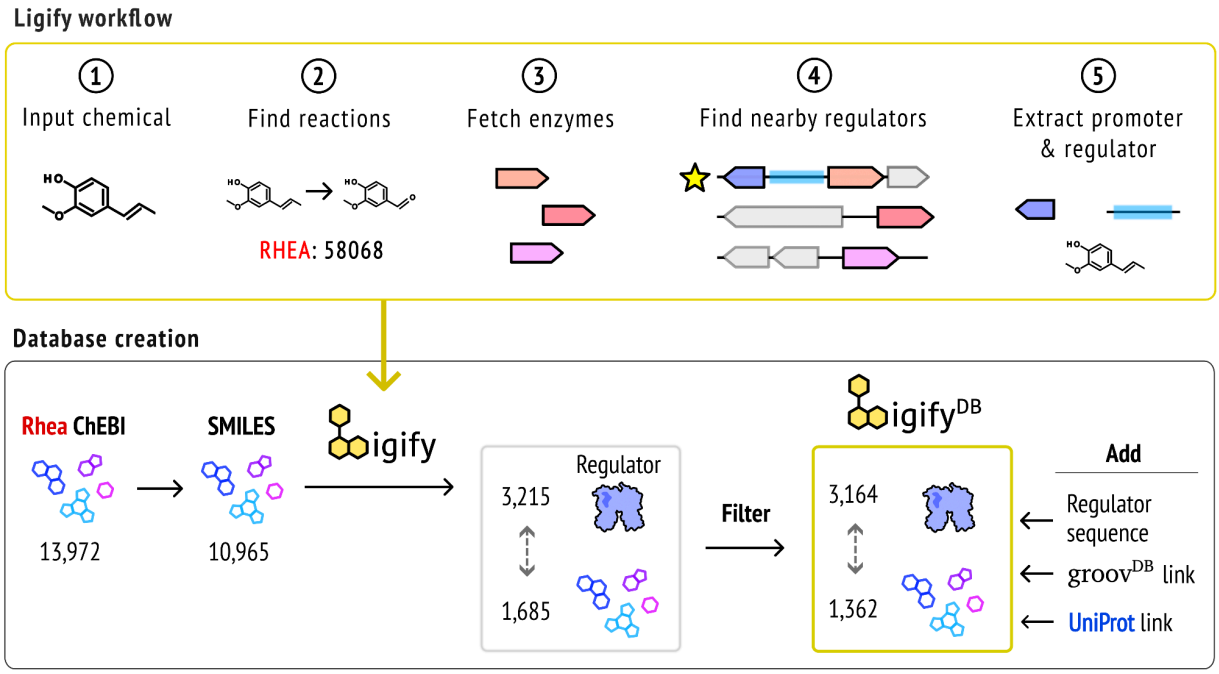


Figure 1: Workflow for creating Ligify^{DB}. (top) The Ligify workflow, which takes a small molecule input and returns transcription factors predicted to interact with the molecule. (bottom) Workflow for creating Ligify^{DB}. Ligify was run on all small molecule SMILES identifiers from Rhea, generating predictions for 1,685 molecules to 3,215 transcription factors. After filtering and data supplementation, the final version of Ligify^{DB} contains 13,435 predicted interactions.

Analysis of database content

To better understand the content of the resulting database, we performed a series of analyses. First we compared the diversity of organisms represented in Ligify^{DB}. Among the 3,164 TF regulators, just over half belong to the Pseudomonadota phyla, with the remaining being represented by 29 other phyla dominated by Actinomycetota and Bacillota (**Figure 2a**). A closer look into the most common genuses reveals a prevalence of model organisms, such as Escherichia and Pseudomonas, as expected (**Figure 2b**). However, over 70% of regulators are not found within the top 10 organisms. Collectively, these results highlight the phylogenetic diversity of regulators within Ligify^{DB}. We next analyzed the composition and distribution of regulator sequences, starting with structural family annotations. The most common structural families include LysR, TetR, GntR, and MarR families, which is expected given the wide representation of these families in nature[21] (**Figure 2c**). Many less common families are also found in the database, such as the RpiR, PyrR, and PmpR structural families. Regulators tend to be between 150 and 350 amino acids long, with peaks around lengths of 180, 230, and 320, which typically correspond to the (MarR/TetR), LysR, and LacI families, respectively (**Figure 2d**). Finally, we compared the diversity of all 1,362 ligands within Ligify^{DB}. Plotting the distribution of chemical structures using molecular fingerprints revealed a wide spread in chemical diversity, with larger clusters around sugars, aromatics, and amino acids, and smaller clusters around Coenzyme-A ligated molecules and cofactor-like molecules (**Figure**

2e). Among these, the most common ligands consisted largely of amino acids and central metabolites like pyruvate, likely because they can serve roles as general cofactors in enzymatic reactions[22] (**Figure 2f**).

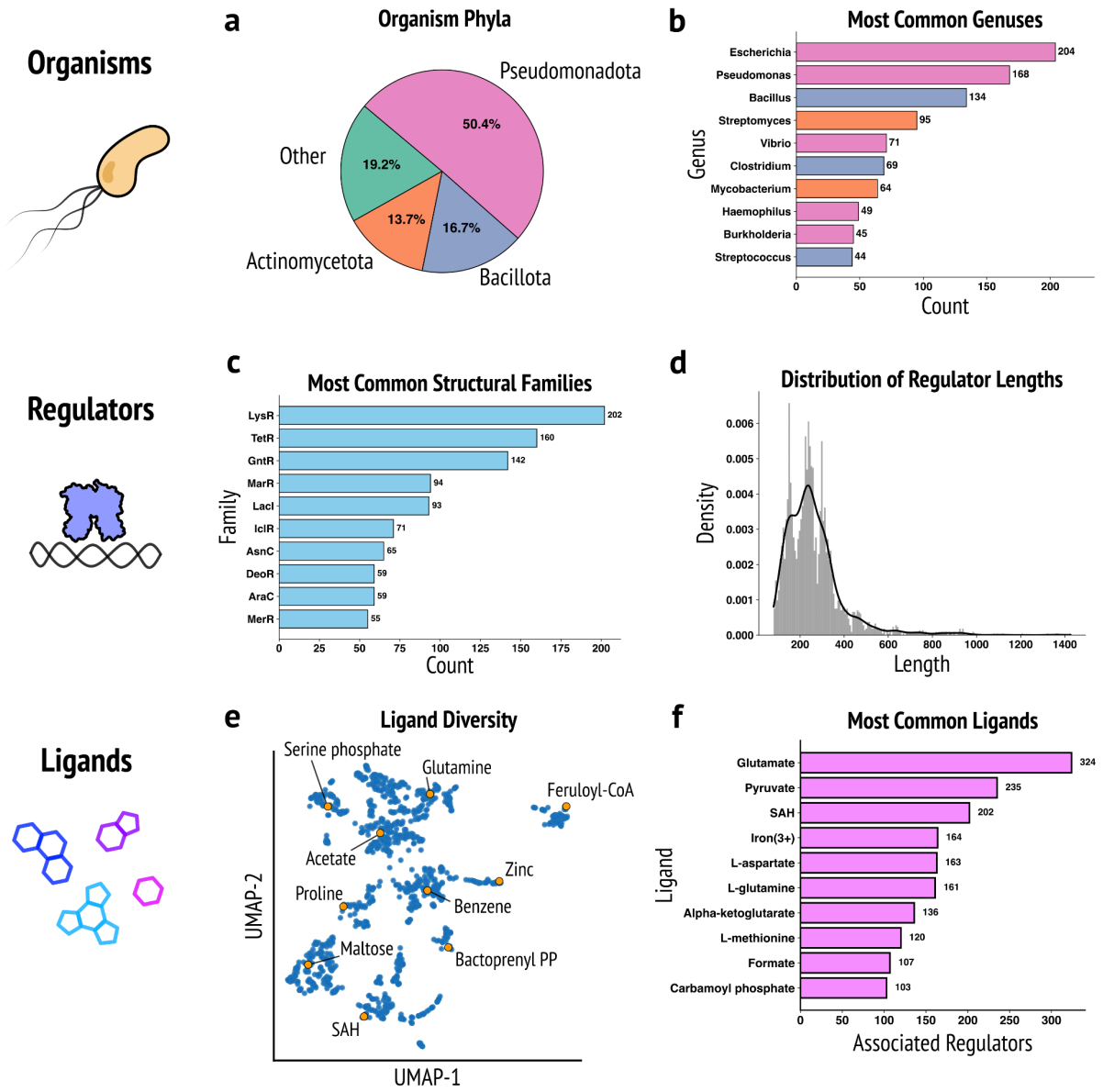


Figure 2: Organism, regulator, and ligand diversity of Ligify^{DB}. (a) Distribution of organism phyla for regulators within the database. (b) The top 10 most common genuses. Bars are colored according to the organism's phyla as shown in panel a. (c) The top 10 most common regulator structural families, indicated from regulator annotations. (d) Distribution of amino acid lengths among the 3,164 regulators in the database. (e) UMAP illustration of the structural diversity among the 1,362 ligands in the database. Highlighted chemicals are labeled and shown as orange dots. All other chemicals are shown as blue dots. (f) The top 10 most common ligands in the database.

A web interface for database search and visualization

The previous version of Ligify was built using Streamlit, which restricted data visualizations and accessibility. To provide an extensible interface where users can easily navigate and interact with Ligify^{DB} content, we created a dedicated web application using the React framework. Users can query the database via a search bar on the home page, which allows for chemical similarity search using a chemical SMILES input, or text search using a regulator's RefSeq ID. Alternatively, users can browse the database from a data table on the "Browse" page, which allows for filtering entries based on sensor attributes, such as length or annotation, or filtering by full-text search.

The new React interface adds several new ways to visualize the database content (**Figure 3**). The 2D structures of all candidate ligands are rendered alongside an interactive AlphaFold model of the TF regulator protein. The protein sequence and predicted promoter sequence are displayed with length indicators. An interactive operon map allows users to study other proximally-encoded genes, which may provide additional insights into the true effector molecule recognized by the regulator. Metadata and literature for the associated enzyme is displayed, which is crucial for evaluating the quality of the prediction. A table containing the rank of the prediction, according to heuristics such as operon length and regulator-enzyme distance, is shown. Finally, a "plasmid builder" tool enables users to design a reporter circuit that uses the predicted regulator to control the expression of a fluorescent protein. The plasmid architecture is designed to be highly insulated and modular by default, incorporating strong terminators before and after each expression unit, and using modular bicistronic RBS elements to tunably control the expression of the regulator protein. After selecting options for genetic elements, such as the fluorescent protein, the promoter and RBS strength driving the regulator, and the plasmid backbone, users can download an annotated genbank file of the designed plasmid.

A detailed documentation page provides context, statistics, and guidance for users. The Ligify prediction model, workflow, and database creation pipeline are all described with illustrations, alongside a detailed comparison of Ligify^{DB} to the original version of Ligify. Descriptions of genetic elements in the "plasmid builder" tool are provided to help users understand which elements to tune and when. A contact form allows users to directly report any issues to the Ligify^{DB} administration team. Finally, an option is available to download the entire 12 MB database, which should be useful for computational biologists who would benefit from programmatic access.

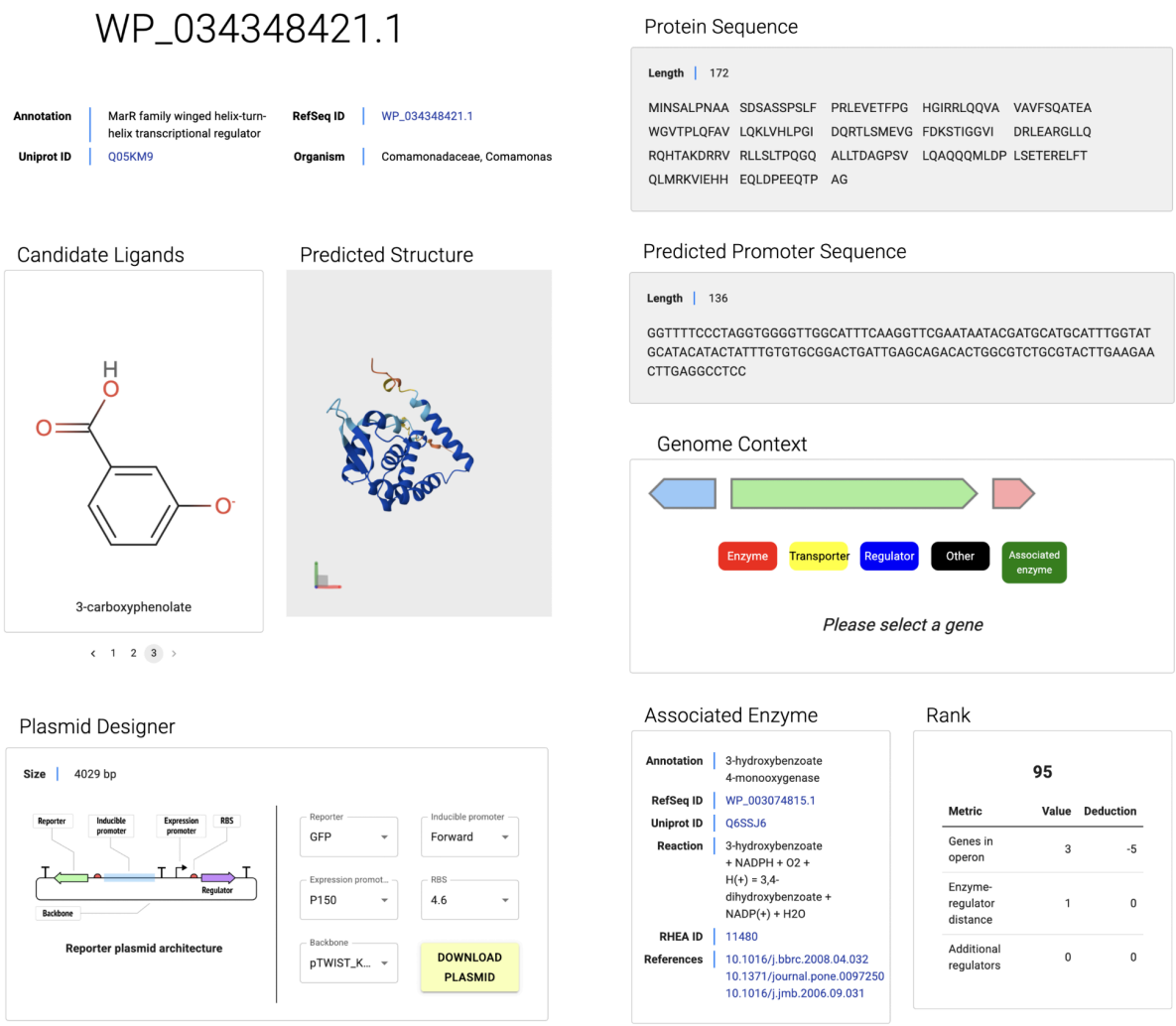


Figure 3: User interface of a representative Ligify^{DB} sensor page. Metadata and links to external databases are provided for each regulator, above their interactive AlphaFold structure and 2D structures of predicted candidate ligands. A plasmid designer tool enables users to create a reporter circuit using the regulator to control the expression of a fluorescent protein. The regulator sequence, predicted promoter sequence, interactive operon model, metadata on the associated enzyme, and calculated rank are also provided in the user interface.

Discussion

Here, we report a platform for predicted TF biosensors, comprising a database of 3,164 TFs associated with 1,362 ligands, and an accompanying interactive web interface for data navigation and visualization. Storing pre-computed prediction results in a static database delivers significant advantages over on-demand prediction tools, like TFBMiner and Ligify[13,14]. Static data is served much faster by requiring less computation, and is easier to maintain by being less reliant on external resources, which would risk losing support over time. Access to all prediction results also enables an in-depth analysis of database

content, as we provide here, and allows for downloading the entire database to facilitate incorporation into bioinformatic workflows.

The content in Ligify^{DB} was found to be relatively diverse, in terms of host phylogeny, TF structural families, and small molecule structures. While there is a bias towards model organisms, even the most represented organism (*E. coli*) comprised ~6% of the database. As expected, predicted ligands were largely composed of primary and secondary natural metabolites. Directed evolution approaches are likely necessary to yield TF biosensors responsive to synthetic molecules[23,24].

We envision this resource as not just a static database, but a dynamic platform that will expand as enzyme reaction databases grow and new prediction methods are developed. For example, 682 new chemicals have been added to the Rhea database since writing this manuscript. Other enzyme databases, such as KEGG and Brenda may also be used to increase the number of predicted TFs[25,26]. Beyond these expert curated databases, enzyme specificity prediction tools may be capable of expanding the scope of enzyme-ligand associations[27]. Separately, biophysical models may also prove useful for predicting TF-ligand interactions where the classic enzyme-association model fails. Machine learning models trained to predict TFs, such as DeepTFactor, can be used alongside AlphaFold to generate a large database of predicted TF structures[28]. Protein-ligand docking software can then be used to suggest possible ligands that fit the binding pocket of each predicted structure[29]. Finally, the resulting TF databases can be supplemented with complementary prediction tools, like Snowprint, to add information about hypothetical DNA binding sequences[30]. While only a tiny fraction of the several million sequenced TF biosensors have been characterized, combining biocuration with computational-guided experimental validation will likely accelerate progress towards mapping all prokaryotic TFs to the ligands they bind.

Materials and Methods

Database curation and analysis

All data processing was performed using Python v3.12.4. The following small molecules failed data fetching or were omitted due to poor likelihood of being true effector molecules (numbers indicate CHEBI IDs): water (15377), hydron (15378), dioxygen (15379), hydrogen peroxide (16240), nitrite (16301), carbon monoxide (17245), sulfite (17359), bicarbonate (17544), nitrate (17632), hydrosulfide (29919), ATP(4-) (30616), dTTP(4−) (37568), hydrogen phosphate (43474), Uridine triphosphase (46398), CoA (57287), FMNH2 (57618), UDP-α-D-glucuronate(3−) (58053). The following molecules were intentionally removed from the database due to their assumed role as cofactors rather than ligand effectors (numbers indicate occurrence): ATP (1102), Phosphate (837), ADP (836), Diphosphate (626), Ammonium (384), APAD (358), Carboxy-SAM (284), Coenzyme A (273), Oxygen (270), Acetyl CoA (209), 5-phosphoribosyl-PP (154), and D-glyceraldehyde 3-phosphate (130). All data is stored both in the public folder of the Ligify^{DB} source code, available on GitHub, as well as on Cloudflare R2, available for download.

Building the Ligify^{DB} Web Application

The frontend web application was developed using React v18.3.1. Material UI (MUI) v7.2.0 was used as a user interface framework. Specific visualization components were built using a set of node libraries. Chemical structures are displayed using smiles-drawer (v2.0.3), interactive protein structures are rendered using nightingale-structure (v5.6.1), and interactive operon diagrams are rendered using react-konva (18.2.10). Data is cached using Zustand (v5.0.8) to reduce subsequent page load times. Chemical similarity search is performed by converting the input SMILES code into a Morgan fingerprint using RDKit (radius: 2, size: 2048) and compared to fingerprints for all ligands within an AWS lambda function. The web application is deployed using AWS-Amplify.

Data availability

All source code of Ligify^{DB} is open-sourced on GitHub under an MIT license and available here: https://github.com/groov-bio/ligify-ui/. The Ligify^{DB} website is publicly accessible here: https://ligify.groov.bio.

Acknowledgements

We are grateful for helpful suggestions from Khalid K. Alam, Adam J. Meyer, and R. C. Baer.

Author contributions: S.D.: Conceptualization, lead software development, and writing. N.Z. and P.T.: Experiments and data analysis. J.D.L. and J. J.: Assisted with software development, and helped revise the original draft. M. S. and P.A.S.: Supervision and writing. S.D.: Conceptualization, software contributions, data curation, and writing.

Ethics declarations

S.D. has financial relationships with Retna Bio LLC. All other authors declare no competing interests.

Funding

Funding is acknowledged from the Harvard Medical School Synthetic Biology HIVE.

References

1. van Sint Fiet, S., van Beilen, J. B. & Witholt, B. Selection of biocatalysts for chemical synthesis. *Proc. Natl. Acad. Sci.* **103**, 1693–1698 (2006).
2. Koberstein, J. N. *et al.* Monitoring glycolytic dynamics in single cells using a fluorescent biosensor for fructose 1,6-bisphosphate. *Proc. Natl. Acad. Sci.* **119**, e2204407119 (2022).
3. Nguyen, P. Q. *et al.* Wearable materials with embedded synthetic biology sensors for biomolecule detection. *Nat. Biotechnol.* **39**, 1366–1374 (2021).
4. Jung, J. K. *et al.* Cell-free biosensors for rapid detection of water contaminants. *Nat. Biotechnol.* **38**, 1451–1459 (2020).
5. Gossen, M. *et al.* Transcriptional Activation by Tetracyclines in Mammalian Cells. *Science* **268**, 1766–1769 (1995).
6. Karp, P. D. *et al.* The BioCyc collection of microbial genomes and metabolic pathways. *Brief. Bioinform.* **20**, 1085–1093 (2019).

7. Parise, M. T. D. *et al.* CoryneRegNet 7, the reference database and analysis platform for corynebacterial gene regulatory networks. *Sci. Data* **7**, 142 (2020).
8. Salgado, H. *et al.* RegulonDB v12.0: a comprehensive resource of transcriptional regulation in E. coli K-12. *Nucleic Acids Res.* **52**, D255–D264 (2024).
9. Ahn, S. K., Cuthbertson, L. & Nodwell, J. R. Genome Context as a Predictive Tool for Identifying Regulatory Targets of the TetR Family Transcriptional Regulators. *PLOS ONE* **7**, e50562 (2012).
10. Cuthbertson, L., Ahn, S. K. & Nodwell, J. R. Deglycosylation as a Mechanism of Inducible Antibiotic Resistance Revealed Using a Global Relational Tree for One-Component Regulators. *Chem. Biol.* **20**, 232–240 (2013).
11. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Res.* **53**, D672–D677 (2025).
12. Bansal, P. *et al.* Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res.* **50**, D693–D700 (2022).
13. Hanko, E. K. R., Joosab Noor Mahomed, T. A., Stoney, R. A. & Breitling, R. TFBMiner: A User-Friendly Command Line Tool for the Rapid Mining of Transcription Factor-Based Biosensors. *ACS Synth. Biol.* **12**, 1497–1507 (2023).
14. d'Oelsnitz, S., Love, J. D., Ellington, A. D. & Ross, D. Ligify: Automated Genome Mining for Ligand-Inducible Transcription Factors. *ACS Synth. Biol.* **13**, 2577–2586 (2024).
15. Koch, M., Pandi, A., Delépine, B. & Faulon, J.-L. A dataset of small molecules triggering transcriptional and translational cellular responses. *Data Brief* **17**, 1374–1378 (2018).
16. Love, J. D. *et al.* groovDB in 2025: A community-editable database of small molecule biosensors. 2025.08.18.670880 Preprint at https://doi.org/10.1101/2025.08.18.670880 (2025).
17. Tellechea-Luzardo, J., Martín Lázaro, H., Moreno López, R. & Carbonell, P. Sensbio: an online server for biosensor design. *BMC Bioinformatics* **24**, 71 (2023).
18. Novichkov, P. S. *et al.* RegPrecise 3.0 – A resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genomics* **14**, 745 (2013).
19. Ashniev, G. A. *et al.* Evolution of transcriptional regulation of histidine metabolism in Gram-positive bacteria. *BMC Genomics* **23**, 558 (2022).
20. Barrows, J. K. & Van Dyke, M. W. A CsoR family transcriptional regulator, TTHA1953, controls the sulfur oxidation pathway in *Thermus thermophilus* HB8. *J. Biol. Chem.* **299**, 104759 (2023).
21. Ortet, P., De Luca, G., Whitworth, D. E. & Barakat, M. P2TF: a comprehensive resource for analysis of prokaryotic transcription factors. *BMC Genomics* **13**, 628 (2012).
22. Kelly, S. A. *et al.* Application of ω-Transaminases in the Pharmaceutical Industry. *Chem. Rev.* **118**, 349–367 (2018).
23. d'Oelsnitz, S. *et al.* Using fungible biosensors to evolve improved alkaloid biosyntheses. *Nat. Chem. Biol.* **18**, 981–989 (2022).
24. Nishikawa, K. K. *et al.* Highly multiplexed design of an allosteric transcription factor to sense new ligands. *Nat. Commun.* **15**, 10001 (2024).
25. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Res.* **53**, D672–D677 (2025).
26. Chang, A. *et al.* BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Res.* **49**, D498–D508 (2021).

27. 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).
28. Kim, G. B., Gao, Y., Palsson, B. O. & Lee, S. Y. DeepTFactor: A deep learning-based tool for the prediction of transcription factors. *Proc. Natl. Acad. Sci.* **118**, e2021171118 (2021).
29. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
30. d'Oelsnitz, S., Stofel, S. K., Love, J. D. & Ellington, A. D. Snowprint: a predictive tool for genetic biosensor discovery. *Commun. Biol.* **7**, 1–9 (2024).