Article

LIGHTHOUSE illuminates therapeutics for a variety of diseases including COVID-19

Hideyuki Shimizu[1,2,3]*, Manabu Kodama[1], Masaki Matsumoto[4], Yasuko Orba[5], Michihito Sasaki[5], Akihiko Sato[5,6], Hirofumi Sawa[5,7,8,9] & Keiichi I. Nakayama[1]*

[1]Department of Molecular and Cellular Biology, Medical Institute of Bioregulation, Kyushu University, Fukuoka, Japan. [2]Department of Systems Biology, Harvard Medical School, Boston, MA, USA. [3]Wyss Institute for Biologically Inspired Engineering, Harvard Medical School, Boston, MA, USA. [4]Department of Omics and Systems Biology, Niigata University Graduate School of Medical and Dental Sciences, Niigata, Japan. [5]Division of Molecular Pathobiology, International Institute for Zoonosis Control, Hokkaido University, Sapporo, Japan. [6]Drug Discovery and Disease Research Laboratory, Shionogi & Co. Ltd., Osaka, Japan. [7]International Collaboration Unit, International Institute for Zoonosis Control, Hokkaido University, Sapporo, Japan. [8]One Health Research Center, Hokkaido University, Sapporo, Japan. [9]Global Virus Network, Baltimore, MD, USA.

* Correspondence: hideyuki_shimizu@hms.harvard.edu (H.S.) or nakayak1@bioreg.kyushu-u.ac.jp (K.I.N.)

SUMMARY

Although numerous promising therapeutic targets for human diseases have been discovered, most have not been successfully translated into clinical practice[1]. A bottleneck in the application of basic research findings to patients is the enormous cost, time, and effort required for high-throughput screening of potential drugs[2] for given therapeutic targets. Recent advances in 3D docking simulations have not solved this problem, given that 3D protein structures with sufficient resolution are not always available and that they are computationally expensive to obtain. Here we have developed LIGHTHOUSE, a graph-based deep learning approach for discovery of the hidden principles underlying the association of small-molecule compounds with target proteins, and we present its validation by identifying potential therapeutic compounds for various human diseases. Without any 3D structural information for proteins or chemicals, LIGHTHOUSE estimates protein-compound scores that incorporate known evolutionary relations and available experimental data. It identified novel therapeutics for cancer, lifestyle-related disease, and bacterial infection. Moreover, LIGHTHOUSE predicted ethoxzolamide as a therapeutic for coronavirus disease 2019 (COVID-19), and this agent was indeed effective against alpha, beta, gamma, and delta variants of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that are rampant worldwide. Given that ethoxzolamide is already approved for several diseases, it could be rapidly deployed for the treatment of patients with COVID-19. We envision that LIGHTHOUSE will bring about a paradigm shift in translational medicine, providing a bridge from bench side to bedside.

INTRODUCTION

Despite enormous efforts to eradicate serious medical conditions such as cancer and infectious diseases, the translation of innovative research results into clinical practice progresses slowly[1], leaving a large gap between bench side and bedside. The difficulty in identifying bioactive chemicals for a given target protein is one reason for this slow progress, with high-throughput screening (HTS) of a sufficiently diverse compound library being required for each target. About 10^{60} natural compounds with a molecular mass of <500 Da are thought to exist[2], but only ~10^6 of these molecules are available for HTS. Over the past few decades, molecular docking simulations have become widely adopted to reduce the cost, time, and effort required for HTS. This approach has been successful for some proteins whose crystal structures have been solved. However, given that high-resolution three-dimensional (3D) structural data are not available for most proteins to date and the high computational requirements of this approach, its application has been limited.

Recent advances in artificial intelligence (AI) have demonstrated its potential in the pharmaceutical industry[3]. Although many AI-based drug discovery methods have been proposed[4–6], they have had limited success in translational medicine, with almost all existing studies having been based solely on *in silico* simulations. In addition, most platforms to date have been trained with small data sets, such as Directory of Useful Decoys Enhanced (DUD-E), that have known biases[7] and are far from reflecting real-world data. Furthermore, many existing methods are based on a single network structure, whereas ensemble learning, which combines multiple network structures with different properties, would be more accurate and appropriate for AI-based drug discovery[8]. As far as we are aware, no published study has described the discovery and validation of therapeutics for multiple human diseases based on the use of a single AI platform.

With this background, we have developed a new AI-based drug discovery platform, designated LIGHTHOUSE (Lead Identification with a GrapH-ensemble network for arbitrary Targets by Harnessing Only Underlying primary SEquence), an ensemble, end-to-end, graph-based deep learning tool that can predict chemicals able to interact with any protein of interest without 3D structural information. We have applied LIGHTHOUSE to malignant, infectious, and metabolic diseases. In addition, we show that LIGHTHOUSE successfully discovered a drug effective against wild-type and variant forms of severe acute

respiratory syndrome coronavirus 2 (SARS-CoV-2), with this drug already having been approved for other purposes. We therefore believe that LIGHTHOUSE will revolutionize drug discovery by identifying, from the vast chemical space, candidate compounds for a given protein with a reduced cost, time, and effort and with a wide range of potential biomedical applications.

RESULTS

LIGHTHOUSE predicts confidence and IC_{50}-related scores for any protein-chemical pair

We developed an end-to-end framework that relies on a message passing neural network (MPNN) for compound embedding[9] to calculate scores for the association between any protein and any chemical. This chemical encoder takes simplified molecular-input line-entry system (SMILES) chemical encoding as input, considers the compounds as (mathematical) graph structures, and transforms them into low-dimensional vector representations. We adopted three different embedding methods for protein sequences: CNN (convolutional neural network)[4], Transformer[10], and AAC (amino acid composition up to 3-mers)[11]. These methods take amino acid sequences and embed them in numerical vectors that take into account nearby (CNN) or distant (Transformer) sequences or physicochemical properties (AAC). The products of these chemical and protein encoding steps are then concatenated and entered into a feed-forward dense neural decoder network. Each chemical-protein pair is converted into a single score after this series of computations (Fig. 1a). We used this architecture to estimate both the confidence level for chemical-protein pairs and their median inhibitory concentration (IC_{50}) values.

To train the platform to estimate confidence, we used ~1.3 million compound–(human) protein interactions (CPIs) stratify-sampled from STITCH (Supplementary Table 1), which is one of the largest CPI databases[12] and registers compound-protein pairs together with confidence scores. These scores are based on experimental data, evolutionary evidence such as homologous protein and compound relations, and co-occurrence frequencies in literature abstracts (scores range from 0 to 1, with 1 being the most reliable). To avoid overfitting, we divided the overall data into training (80%), validation (10%), and test (10%) data sets (Extended Data Fig. 1a).

We fed the network with protein primary structures and chemicals and trained it to output the scores from the STITCH training data set (Fig. 1b). When

we trained the three models (CNN, AAC, and Transformer for protein encoders) separately, the mean squared error (MSE) for the validation data was gradually decreased, and the area under the receiver operating characteristic curve (AUROC) was also improved (Extended Data Fig. 1b–g). These findings indicated that our AI models correctly learned the hidden 1D relation underlying the compound-protein pairs without overfitting the training data. We examined the performance of the models with the test data set at the end of the training and (epoch-wise) validation phases, and we discovered that the AUROC for all three models was >0.80 (Supplementary Table 2). These scores are equivalent to or better than those of cutting-edge 3D docking simulations[13-15]. It is of note that our AI models can be applied to proteins for which 3D structural information is not available. We took the harmonic mean of the three scores to define the confidence score (Fig. 1b).

We also trained the models to predict scores based on IC_{50} values. For this purpose, we used data from BindingDB[16], which collects a variety of experimental findings, and we divided the data into training (80%), validation (10%), and test (10%) data sets (Extended Data Fig. 2a). The same architecture was adopted to train the AI models to predict scaled IC_{50} values (Fig. 1c), yielding an interaction score, and we confirmed that the models adequately learned how to predict IC_{50} from amino acid sequence–chemical pairs (Extended Data Fig. 2b-g). Finally, we assessed the performance of the models with undisclosed test data, finding that they performed well in predicting IC_{50} (Supplementary Table 3).

In silico verification of LIGHTHOUSE

We next evaluated the performance of LIGHTHOUSE in terms of its ability to predict known CPIs. We generated two data sets for this purpose: a “Positive” data set consisting of reliable CPIs (STITCH confidence score of >0.9), and a “Negative” data set in which the amino acid sequences of the “Positive” data set were inverted so that they would no longer be expected to interact with the corresponding chemicals. Calculation by LIGHTHOUSE of the confidence scores for both data sets revealed that those for the "Positive" data set were heavily skewed toward 1 (Fig. 2a). Receiver operating characteristic (ROC) curve analysis showed that the two data sets could be distinguished on the basis of their LIGHTHOUSE confidence scores (Fig. 2b). Given that the STITCH database used for the training of LIGHTHOUSE relies not only on

experimental CPI data but also on co-appearance of chemicals and proteins in the literature, some well-studied molecules, such as ATP, have high values even in the “Negative" data set. Despite the presence of such false positives, LIGHTHOUSE proved to be effective in predicting the degree of association between protein-chemical pairs solely on the basis of protein primary structure.

We next validated the effectiveness of LIGHTHOUSE for well-studied compound-protein pairs. LIGHTHOUSE yielded high confidence scores for adrenergic receptors (α1, α2, β1, β2, and β3) and epinephrine (Fig. 2c). Histamine receptors are classified into four subtypes[17], with HRH1 and HRH2 being targets of antiallergy and antiulcer drugs, respectively. LIGHTHOUSE predicted that the HRH1 antagonist fexofenadine would associate to a greater extent with HRH1 than with HRH2, whereas the HRH2 inhibitor famotidine would associate to a greater extent with HRH2 than with HRH1 (Fig. 2d). These results suggested that LIGHTHOUSE is able to accurately discriminate receptor subtype–level differences solely on the basis of amino acid sequences.

LIGHTHOUSE also proved informative both for macrocyclic chemicals such as rapamycin, yielding a high confidence score for this drug and mechanistic target of rapamycin (MTOR) (Fig. 2e), as well as for peptide drugs such as bortezomib (used for treatment of multiple myeloma), leuprorelin (hormone-responsive cancers), and semaglutide (type 2 diabetes) (Fig. 2f), yielding high confidence scores for these drugs and their known targets: proteasome subunit PSMB1[18], gonadotropin-releasing hormone receptor (GNRHR)[19], and glucagon-like peptide–1 (GLP-1) receptor (GLP1R)[20], respectively. Given the rapidly growing demand for peptide drugs[21], LIGHTHOUSE will prove useful for the development of novel peptide therapeutics for a variety of promising targets.

We also applied LIGHTHOUSE to five drugs that were approved by the U.S. Food and Drug Administration (FDA) in 2020 but which had not yet been registered in the STITCH database. LIGHTHOUSE successfully predicted the association between these new drugs and their target proteins (Fig. 2g), indicating the expandability of LIGHTHOUSE to a much larger exploration space than that encompassed by STITCH. This series of findings thus demonstrated the ability of LIGHTHOUSE to discover new drugs for a broad spectrum of diseases.

LIGHTHOUSE discovers an inhibitor of PPAT, a key metabolic enzyme for cancer treatment

We investigated whether LIGHTHOUSE can identify compounds for novel and potentially important therapeutic targets. As such a target, we chose phosphoribosyl pyrophosphate amidotransferase (PPAT), a rate-limiting enzyme in the *de novo* nucleotide synthesis pathway, given that its expression is most correlated among all metabolic enzymes with poor prognosis in various human cancers and that its depletion markedly inhibits tumor growth[22]. Although no PPAT inhibitor has been developed and the 3D structure of the protein has not been solved, we attempted to discover an inhibitor for PPAT by LIGHTHOUSE solely on the basis of its amino acid sequence. We virtually screened ~10^9 commercially available compounds in the ZINC database[23] (Extended Data Fig. 3). To reduce the calculation time, we adopted a step-by-step application of LIGHTHOUSE (Fig. 3a). The MPNN_CNN model excluded most of the chemicals unrelated to PPAT, with only 2.41% of the starting compounds having a score of >0.5 in this initial screening (Fig. 3b). The selected compounds were then processed by the MPNN_AAC and MPNN_Transformer models, which reduced the number of candidate chemicals to 0.0356% of the initial compounds. We also calculated interaction scores by LIGHTHOUSE and visualized them in a 2D plot (Fig. 3c, left). The best candidates would be expected to have high confidence and interaction scores, appearing in the upper right corner of the plot. Indeed, this criterion was met by several well-known drug-target combinations (Fig. 3c, right).

The top candidate for a PPAT inhibitor in terms of confidence score was ZINC8551105 (riboflavin 5′-monophosphate), with a predicted IC_{50} of 1 to 10 μM (Fig. 3d). We performed a biochemical assay to test this prediction and found that riboflavin 5′-monophosphate indeed markedly inhibited PPAT activity with an actual IC_{50} of 7 μM (Fig. 3e). This compound, discovered by LIGHTHOUSE solely on the basis of the PPAT amino acid sequence, is thus a potential lead compound for the development of new therapeutics targeted to a variety of cancers.

LIGHTHOUSE identifies an inhibitor of drug-resistant bacterial growth

Bacterial infections pose a clinical problem worldwide, especially in developing countries, and the emergence of drug-resistant bacterial strains as a result of the overuse of antibiotics has exacerbated this problem. β-Lactamase enzymes

produced by antibiotic-resistant bacteria[24] target the β-lactam ring of antibiotics of the penicillin family. We therefore applied LIGHTHOUSE to search for antibiotics not dependent on β-lactam structure.

LIGHTHOUSE predicted that pyridoxal 5'-phosphate might associate with penicillin binding proteins such as PBP2 (*mrdA*), PBP3 (*ftsL*), and PBP5 (*dacA*), all of which are essential for cell wall synthesis in *Escherichia coli*[25] (Fig. 3f). This compound indeed suppressed the growth of *E. coli* strain JM109 in a concentration-dependent manner (Extended Data Fig. 4). Importantly, pyridoxal 5'-phosphate also markedly inhibited the growth of an ampicillin-resistant *E. coli* transformant that produces β-lactamase (Fig. 3g). These results thus suggested that, even though it was trained with human proteins, LIGHTHOUSE can also be applied to nonhuman (even bacterial) proteins.

LIGHTHOUSE informs optimization of lead compounds

Diabetes mellitus is also a serious public health concern, with the number of affected individuals expected to increase markedly in the coming decades[26]. Dipeptidyl peptidase–4 (DPP-4) cleaves and inactivates the incretin hormones GLP-1 and glucose-dependent insulinotropic polypeptide (GIP), and DPP-4 inhibitors are a new class of antidiabetes drug[27]. Given that LIGHTHOUSE also predicts interaction scores, we examined whether it might also contribute to the optimization step of drug development. Indeed, LIGHTHOUSE accurately predicted the rank order of potency for several recently identified DPP-4 inhibitor derivatives[28] (Extended Data Fig. 5a). Furthermore, LIGHTHOUSE predicted that removal of the phosphate group would reduce the inhibitory potency of riboflavin 5′-monophosphate for PPAT (Fig. 3e), and this prediction was confirmed correct by the finding that the IC_{50} value was increased from 7 to 49.9 μM (Extended Data Fig. 5b).

LIGHTHOUSE is also able to estimate the effect of point mutations on CPIs. For example, the T315I mutation of ABL1 in leukemia cells reduces the efficacy of imatinib[29], and LIGHTHOUSE accurately predicted the effect of this mutation (Extended Data Fig. 5c). LIGHTHOUSE is able to provide such insight from only wild-type amino acid sequences, given the lack of variant information in the original training data set. Our results suggest that LIGHTHOUSE is able to predict the effects of small changes in protein or chemical structure, and that this will be the case even if such variants do not exist in nature.

LIGHTHOUSE identifies potential on- and off-targets of given compounds

Opposite to the mode of drug discovery for a given protein, LIGHTHOUSE should also be able to identify proteins as potential on- or off-targets for a given compound. To verify this notion, we examined statins, which are HMG-CoA reductase inhibitors widely administered for the treatment of hyperlipidemia. Epidemiological studies have shown that statins not only lower cholesterol, however, but also have effects on cancer, although the target molecules for these effects have remained unclear[30]. We therefore applied LIGHTHOUSE to three representative statins (atorvastatin, cerivastatin, and fluvastatin) and computed confidence scores for all human protein-coding genes (Fig. 4a, Extended Data Fig. 6a, Supplementary Table 4). We then sorted the genes on the basis of these confidence scores and performed Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis for the top 500 potential statin targets. In addition to lipid-related pathways such as atherosclerosis and fatty liver, “pathways in cancer” was one of the most enriched KEGG pathways (Extended Data Fig. 6b), consistent with previous findings[31–34]. Potential targets of statins for cancer treatment identified by LIGHTHOUSE included STAT3, CCND1, AKT1, and CCL2 (Extended Data Fig. 6c).

Given that side effects of drugs often manifest in organs that express target proteins, we hypothesized that LIGHTHOUSE might be able to identify which organs are at risk of damage from a given drug. We performed another enrichment analysis for the same top 500 potential statin target genes to determine which organs or cell types preferentially express these genes. The top three candidates were the liver, adipocytes, and lung (Extended Data Fig. 6d), consistent with the liver being the primary site of statin metabolism and interstitial pneumonia being one of the most severe side effects of statins[35]. Prediction of potential target proteins for a given drug by LIGHTHOUSE will thus provide insight into which organs warrant close monitoring by physicians during treatment with the drug, especially in first-in-human clinical trials.

LIGHTHOUSE identifies novel potential therapeutics for COVID-19

SARS-CoV-2 emerged at the end of 2019 and has caused a pandemic of infectious pulmonary disease, COVID-19[36]. We noticed that genes whose expression is up-regulated after SARS-CoV-2 infection[37–39] were enriched in the list of potential statin targets identified by LIGHTHOUSE (Fig. 4b). Indeed,

previous studies have shown that statins prevent exacerbation of COVID-19[40,41]. With this finding that LIGHTHOUSE is also effective for COVID-19 drug discovery, we applied it to the virtual screening of ~10,000 approved drugs, given that drug repurposing may allow faster delivery of effective agents to patients in need. We calculated scores for angiotensin-converting enzyme 2 (ACE2), which is targeted by SARS-CoV-2 for infection of host cells[42], and the top drug candidate, ethoxzolamide, was selected for validation analysis (Fig. 4c). Immunocytofluorescence analysis revealed that ethoxzolamide blocks proliferation of SARS-CoV-2 in Vero-TMPRSS2 cells (Extended Data Fig. 7). Furthermore, ethoxzolamide was effective against not only the wild-type (Wuhan) virus but also the alpha (U.K.), beta (South Africa), gamma (Brazil), and delta (India) variants. It thus rescued virus-challenged cells in a concentration-dependent manner without affecting noninfected cells (median cytotoxicity concentration > 50 μM) (Extended Data Fig. 8, Supplementary Table 5), and it reduced the virus load present in the culture supernatant of the cells (Fig. 4d, e; Extended Data Fig. 9). Ethoxzolamide is approved for the treatment of seizures and glaucoma[43,44], and its pharmacodynamics are therefore known. It is therefore immediately available for repurposing for the treatment of patients with COVID-19, with its further optimization having the potential to save many lives.

DISCUSSION

Although recent advances in biological and medical research have uncovered various proteins as promising therapeutic targets in a variety of diseases, the clinical application of these research findings has been limited because of the difficulty in identifying therapeutic chemicals for these targets in a cost-effective and high-throughput manner. Acquisition of 3D structural data for target proteins has been labor-intensive, and processing of such data requires a huge amount of computer capacity and time, resulting in a delay in the translation of research findings from the laboratory to the clinic. We have now shown that LIGHTHOUSE facilitates the identification, from a vast chemical space, of candidate compounds for given target proteins solely on the basis of the primary structure of these proteins. Furthermore, the AUROC for LIGHTHOUSE is equivalent to or better than that for state-of-the-art 3D docking simulation methods.

We have applied LIGHTHOUSE to attractive targets for various diseases, including cancer, bacterial infection, metabolic diseases, and COVID-19, with some of the suggested chemicals being determined experimentally to be effective for inhibition of the corresponding targets. LIGHTHOUSE can be applied not only for the identification of lead compounds but also for their subsequent optimization, which requires extensive work to identify more potent and specific or less toxic derivatives. One promising method to support such optimization is to apply LIGHTHOUSE and either reinforcement learning[45] (Extended Data Fig. 10) or Metropolis-Hasting (MH) approaches together. Virtual libraries can be generated from identified lead compounds in an intensive manner with the use of sophisticated chemoinformatics algorithms such as RECAP (Retrosynthetic Combinatorial Analysis Procedure)[46]. Given the recent success of the MH approach in various life science fields[47,48], LIGHTHOUSE should also facilitate optimization of drug candidates.

A limitation of LIGHTHOUSE is the generation of false positives, which is due in part to the fact that the confidence score provided by STITCH is not based solely on experimental data but also on other factors such as co-occurrence in the literature. Well-studied molecules are thus prone to score higher than others. This drawback can be mitigated partially by combining the three different models (CNN, AAC, and Transformer). It may also be important to perform a counter–virtual screening to determine whether an identified small molecule reacts specifically with the target protein or whether it scores highly with many proteins. Such an approach has the potential to reduce the number of false positives and provide more accurate guidance.

Despite this limitation, LIGHTHOUSE proved to be effective for the identification of lead compounds for all conditions tested. It can theoretically be applied to any protein of any organism, and even to proteins that do not exist naturally. This is an advantage over 3D docking simulation methods, which require prior 3D structural knowledge of the protein of interest. LIGHTHOUSE computes and embeds structural information in numerical vectors, which are then readily retrieved by the subsequent decoding module. Given the accelerating development of protein embedding technologies[49] and graph-based chemoinformatics approaches, LIGHTHOUSE has the potential to be a cornerstone of drug discovery.

In summary, we have developed LIGHTHOUSE as a means to discover promising lead compounds for any target protein irrespective of its 3D structural

information. Furthermore, we have demonstrated the power of LIGHTHOUSE by identifying and validating novel therapeutics for various global health concerns including COVID-19. LIGHTHOUSE will serve as a guide for researchers in all areas of biomedicine, paving the way for a wide range of future applications.

REFERENCES

1. Klein, M.E., Parvez, M.M. & Shin, J.-G. Clinical implementation of pharmacogenomics for personalized precision medicine: barriers and solutions. *J. Pharm. Sci.* **106**, 2368–2379 (2017).

2. Dobson, C. M. Chemical space and biology. *Nature* **432**, 824–828 (2004).

3. Paul, D. *et al*. Artificial intelligence in drug discovery and development. *Drug Discov. Today* **26**, 80–93 (2021).

4. Öztürk, H., Özgür A. & Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction. *Bioinformatics* **34**, i821–i829 (2018).

5. Huang, K. *et al*. DeepPurpose: a deep learning library for drug-target interaction prediction. *Bioinformatics* **36**, 5545–5547 (2021).

6. Tsubaki, M., Tomii, K. & Sese, J. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* **35**, 309–318 (2019).

7. Chen, L. *et al*. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS One* **14**, e0220113 (2019).

8. Hansen, L. K. & Salamon, P. Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.* **12**, 993–1001 (1990).

9. Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O. & Dahl, G.E. Neural message passing for quantum chemistry. *arXiv* 1704.01212 [cs.LG] (2017).

10. Vaswani, A. *et al*. Attention is all you need. *arXiv* 1706.03762 [cs.CL] (2017).

11. Reczko, M. & Bohr, H. The DEF data base of sequence based protein fold class predictions. *Nucleic Acids Res.* **17**, 3616–3619 (1994).

12. Szklarczyk, D. *et al*. STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.* **44**, D380–D384 (2016).

13. Hsin, K. *et al*. systemsDock: a web server for network pharmacology-based prediction and analysis. *Nucleic Acids Res.* **44**, W507–W513 (2016).

14. Wang, S., Jiang, J.-H., Li, R.-Y. & Deng, P. Docking-based virtual screening of TβR1 inhibitors: evaluation of pose prediction and scoring functions. *BMC Chem.* **14**, 52 (2020).

15. Moussa, N., Hassan, A. & Gharaghani, S. Pharmacophore model, docking, QSAR, and molecular dynamics simulation studies of substituted cyclic imides and herbal medicines as COX-2 inhibitors. *Heliyon* **7**, e06605 (2021).
16. Gilson, M. K. *et al*. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **44**, D1045–D1063 (2016).
17. Seifert, R. *et al*. Molecular and cellular analysis of human histamine receptor subtypes. *Trends Pharmacol. Sci.* **34**, 33–58 (2013).
18. Berkers, C. *et al*. Activity probe for in vivo profiling of the specificity of proteasome inhibitor bortezomib. *Nat. Methods* **2**, 357–362 (2005).
19. Borroni, R. *et al*. Expression of GnRH receptor gene in human ectopic endometrial cells and inhibition of their proliferation by leuprolide acetate. *Mol. Cell. Endocrinol.* **159**, 37–43 (2000).
20. Knudsen, L.B. & Lau, J. The discovery and development of liraglutide and semaglutide. *Front. Endocrinol. (Lausanne)* **10**, 155 (2019).
21. Muttenthaler, M., King, G.F., Adams, D.J. & Alewood, P.F. Trends in peptide drug discovery. *Nat. Rev. Drug Discov.* **20**, 309–325 (2021).
22. Kodama, M. *et al*. A shift in glutamine nitrogen metabolism contributes to the malignant progression of cancer. *Nat. Commun.* **11**, 1320 (2020).
23. Sterling, T. & Irwin, J.J. ZINC 15—ligand discovery for everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).
24. Tooke, C. L. *et al*. β-Lactamases and β-lactamase inhibitors in the 21st century. *J. Mol. Biol.* **431**, 3472–3500 (2019).
25. Macheboeuf, P., Contreras-Martel, C., Job, C., Dideberg, O. & Dessen, A. Penicillin binding proteins: key players in bacterial cell cycle and drug resistance processes. *FEMS Microbiol. Rev.* **30**, 673–691 (2006).
26. Zheng, Y., Ley, S.H. & Hu, F.B. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nat. Rev. Endocrinol.* **14**, 88–98 (2018).
27. Deacon, C. F. Dipeptidyl peptidase 4 inhibitors in the treatment of type 2 diabetes mellitus. *Nat. Rev. Endocrinol.* **16**, 642–653 (2020).
28. Li, Q., Han, L., Zhang, B., Zhou, J. & Zhang, H. Synthesis and biological evaluation of triazole based uracil derivatives as novel DPP-4 inhibitors. *Org. Biomol. Chem.* **14**, 9598–9611 (2016).

29. Jabbour, E. *et al*. Characteristics and outcomes of patients with chronic myeloid leukemia and T315I mutation following failure of imatinib mesylate therapy. *Blood* **112**, 53–55 (2008).
30. Mei, Z. *et al*. Effects of statins on cancer mortality and progression: a systematic review and meta-analysis of 95 cohorts including 1,111,407 individuals. *Int. J. Cancer* **140**, 1068–1081 (2017).
31. Matusewicz, L., Meissner, J., Toporkiewicz, M. & Sikorski, A.F. The effect of statins on cancer cells—review. *Tumour Biol.* **36**, 4889–4904 (2015).
32. Ahern, T.P., Lash, T.L., Damkier, P., Christiansen, P.M. & Cronin-Fenton, D.P. Statins and breast cancer prognosis: evidence and opportunities. *Lancet Oncol.* **15**, e461–e468 (2014).
33. Mullen, P.J., Yu, R., Longo, J., Archer, M.C. & Penn, L.Z. The interplay between cell signalling and the mevalonate pathway in cancer. *Nat. Rev. Cancer* **16**, 718–731 (2016).
34. Alfaqih, M.A., Allott, E.H., Hamilton, R.J., Freeman, M.R. & Freedland, S.J. The current evidence on statin use and prostate cancer prevention: Are we there yet? *Nat. Rev. Urol.* **14**, 107–119 (2017).
35. Momo, K., Takagi, A., Miyaji, A. & Koinuma, M. Assessment of statin-induced interstitial pneumonia in patients treated for hyperlipidemia using a health insurance claims database in Japan. *Pulm. Pharmacol. Ther.* **50**, 88–92 (2018).
36. Hu, B., Guo, H., Zhou, P. & Shi, Z.-L. Characteristics of SARS-CoV-2 and COVID-19. *Nat. Rev. Microbiol.* **19**, 141–154 (2021).
37. Blanco-Melo, D. *et al*. Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell* **181**, 1036–1045 (2020).
38. Riva, L. *et al*. Discovery of SARS-CoV-2 antiviral drugs through large-scale compound repurposing. *Nature* **586**, 113–119 (2020).
39. Wyler, E. *et al*. Transcriptomic profiling of SARS-CoV-2 infected human cell lines identifies HSP90 as target for COVID-19 therapy. *iScience* **24**, 102151 (2021).
40. Zhang, X.-J. *et al*. In-hospital use of statins is associated with a reduced risk of mortality among individuals with COVID-19. *Cell Metab.* **32**, 176–187 (2020).
41. Gupta, A. *et al*. Association between antecedent statin use and decreased mortality in hospitalized patients with COVID-19. *Nat. Commun.* **12**, 1325 (2021).

42. Walls A. C. *et al*. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* **181**, 281–292 (2020).

43. Pospelov, A.S., Ala-Kurikka, T., Kurki, S., Voipio, J. & Kaila, K. Carbonic anhydrase inhibitors suppress seizures in a rat model of birth asphyxia. *Epilepsia* (in press) doi: 10.1111/epi.16963.

44. Ghorai, S. *et al*. Structure-activity relationship of human carbonic anhydrase-II inhibitors: detailed insight for future development as anti-glaucoma agents. *Bioorg. Chem.* **95**, 103557 (2020).

45. Pereira, T., Abbasi, M., Ribeiro, B. & Arrais, J.P. Diversity oriented deep reinforcement learning for targeted molecule generation. *J. Cheminform.* **13**, 21 (2021).

46. Lewell, X.Q., Judd, D.B., Watson, S.P. & Hann, M.M. RECAP—retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **38**, 511–522 (1998).

47. Biswas, S., Khimulya, G., Alley, E.C., Esvelt, K.M. & Church, G.M. Low-N protein engineering with data-efficient deep learning. *Nat. Methods* **18**, 389–396 (2021).

48. Chen, X. *et al*. Tissue-specific enhancer functional networks for associating distal regulatory regions to disease. *Cell Syst*. **12**, 353–362 (2021).

49. Bepler, T. & Berger, B. Learning the protein language: evolution, structure, and function. *Cell Syst*. **12**, 654–669 (2021).

FIGURE LEGENDS

Fig. 1 Development of LIGHTHOUSE for discovery of drug candidates without 3D structural data. a, The basic network structure of LIGHTHOUSE consists of encoder and decoder networks. The basic network encodes the amino acid sequence of the protein of interest as numerical vectors by one of three independent methods: CNN, AAC, and Transformer. It also takes the SMILES representation of each small-molecule compound and computes the neural representation with the MPNN algorithm. The network then concatenates the protein and compound representations and calculates a "Score." **b**, **c**, LIGHTHOUSE consists of two modules. Module 1 estimates the association between a given compound-protein pair, and module 2 predicts a scaled IC_{50} value for the pair. In each module, the three different streams of the basic network (MPNN_CNN, MPNN_AAC, and MPNN_Transformer) are used, and the harmonic mean of the three scores is presented as the final ensemble score. Each of the three streams in module 1 (**b**) is trained to minimize the error between the predicted "Score" and the score registered in the STITCH database, which contains millions of known and estimated CPIs. The higher the confidence score (closer to 1), the more confident LIGHTHOUSE is that there is some relation between the compound and the protein; conversely, the lower the confidence score (closer to 0), the more confident LIGHTHOUSE is that there is no such relation. Each of the three streams in module 2 (**c**) is trained to predict scaled IC_{50} values with the use of BindingDB data. For instance, an interaction score of 4 means that, if the compound has inhibitory activity, the IC_{50} would be ~100 μM, whereas an interaction score of 9 means that, if the compound has inhibitory activity, the IC_{50} would be ~1 nM. Note that module 2 only works if the compound and protein interact, so this module is auxiliary to module 1.

Fig. 2 *In silico* verification of LIGHTHOUSE. a, For investigation of whether LIGHTHOUSE is able to enrich for compounds with known targets, two data sets were generated from STITCH: a "Positive" data set consisting of CPIs with high scores (>0.9), and a "Negative" data set consisting of the same CPIs but with the amino acid sequences of the proteins reversed (for example, MTSAVM to MVASTM). Proteins in the "Negative" data set would not be expected to interact with the corresponding compounds. LIGHTHOUSE tended to yield higher confidence scores for CPIs in the "Positive" data set, with the exception of the rightmost peak for the "Negative" data set, presumably because these

chemicals (such as ATP) are well known and frequently mentioned in the PubMed literature. **b**, ROC curve showing that LIGHTHOUSE was able to distinguish the "Positive" and "Negative" data sets. **c**–**f**, Known CPIs and their confidence scores predicted by LIGHTHOUSE. **c**, Epinephrine and α-adrenergic (ADRA) and β-adrenergic (ADRB) receptors. **d**, Fexofenadine and the histamine receptor HRH1, and famotidine and the histamine receptor HRH2. **e**, The macrocyclic drug rapamycin and MTOR. **f**, The peptide drugs bortezomib, leuprorelin, and semaglutide and their targets PSMB1, GNRHR, and GLP1R, respectively. **g**, Application of LIGHTHOUSE to five drugs approved by the FDA in 2020 that were not included in the training data set (published in 2016). FNTA, protein farnesyltransferase/geranylgeranyltransferase type–1 subunit α; COMT, catechol O-methyltransferase; S1PR1, sphingosine 1-phosphate receptor 1; DRD2, D2 dopamine receptor.

Fig. 3 Discovery of lead compounds for treatment of cancer or bacterial infection. a, Scheme for PPAT inhibitor discovery. The amino acid sequence of PPAT (517 residues) and the SMILE representation for each chemical were entered into the MPNN_CNN model. If the predicted score was >0.5, the compound was entered into MPNN_AAC, and if the new predicted score was >0.5, the compound was entered into MPNN_Transformer. The harmonic mean of the three scores was then computed to obtain the confidence score. **b**, Almost 1 billion compounds in the ZINC database were processed as in **a**. The first filter (MPNN_CNN score > 0.5) and subsequent two filters (MPNN_AAC score > 0.5, MPNN_Transformer score > 0.5) greatly reduced the initial chemical space (to 0.0356%). The interaction scores for these selected candidates were then also calculated. **c**, A 2D map of the 333,290 selected candidates from **b** is shown on the left. Ideal candidates would be expected to have high confidence and interaction scores and would be plotted in the upper right corner of the map. Indeed, some well-known drug-target pairs meet this criterion, as shown on the right, with compounds represented by the blue circles in the shaded area potentially possessing inhibitory activity for PPAT. ABL1, ABL proto-oncogene 1; COX1, cyclooxygenase 1; HMG-CoA, 3-hydroxy-3-methylglutaryl–coenzyme A. **d**, The top hit compound, ZINC8551105 (riboflavin 5′-monophosphate), is shown together with its confidence score and estimated IC_{50} value. **e**, *In vitro* PPAT activity assay performed in the presence

of various concentrations of riboflavin 5′-monophosphate, with the determined IC_{50} value being within the range predicted by LIGHTHOUSE. Data are shown for four biological replicates. **f**, LIGHTHOUSE predicted that pyridoxal 5′-phosphate would associate with several penicillin binding proteins of *E. coli* (strain K12). mrdA and ftsI are peptidoglycan D,D-transpeptidases, whereas dacA is a D-alanyl-D-alanine carboxypeptidase. **g**, The JM109 strain of *E. coli* was transformed with the pBlueScript II SK+ plasmid, which contains an ampicillin resistance gene as a selection marker, and the cells were plated on LB agar plates containing ampicillin in the absence or presence of pyridoxal 5′-phosphate (3 mg/ml) and were incubated overnight.

Fig. 4 LIGHTHOUSE-based drug repurposing. a, Identification of statin targets by LIGHTHOUSE. LIGHTHOUSE was applied to calculate confidence scores for all human protein-coding genes in the UniProt database and fluvastatin, atorvastatin, and cerivastatin. The harmonic mean of these confidence scores (FluvastatinScore, AtorvastatinScore, and CerivastatinScore) was calculated as an affinity score for statins. Sorting on the basis of this affinity score yielded a list of potential statin target proteins. HMGCR (HMG-CoA reductase), a known key target of statins, was ranked 136th with a score of 0.790. The top 500 identified genes were then subjected to enrichment analysis. LDLR, low-density lipoprotein receptor; APOE, apolipoprotein E; SCD, stearoyl-CoA desaturase; STAT3, signal transducer and activator of transcription 3. **b**, Enrichment analysis of the top 500 potential statin targets for COVID-19–associated gene sets. Minus $\log_{10}$-transformed q values are shown. **c**, Prediction by LIGHTHOUSE of ethoxzolamide as a potential therapeutic for SARS-CoV-2 infection on the basis of its confidence and interaction scores for ACE2. **d**, **e**, Effect of ethoxzolamide on the SARS-CoV-2 load in culture supernatants of Vero-TMPRSS2 cells challenged with Wuhan (WK-521) or India (delta) strains of the virus, respectively. Data are from four independent experiments, with the graph line connecting mean values. $TCID_{50}$, median tissue culture infectious dose; N.D., not detected.

METHODS

Generation of a data set for the training phase of LIGHTHOUSE

The compound SMILES strings of the data set were extracted from the PubChem compound database on the basis of compound names and PubChem compound IDs (CIDs). The protein sequences of the data set were extracted from the UniProt protein database on the basis of gene names/RefSeq accession numbers or the UniProt IDs. We downloaded the protein-chemical link data set of *Homo sapiens* (Taxonomy ID 9606) from the STITCH database (version 5.0). Given that the STITCH score is heavily biased toward 0, we separated the data into nine bins on the basis of the score and stratify-extracted the same number of CPIs (140,000 each), yielding 1,260,000 CPIs (Supplementary Table 1). We then randomly separated these data into training (80%), validation (10%), and test (10%) data sets (Extended Data Fig. 1a). With regard to IC_{50}, we downloaded data from BindingDB, obtained SMILES expressions and amino acid sequences similarly, and again separated the data into training (80%), validation (10%), and test (10%) data sets (Extended Data Fig. 2a). Given that IC_{50} values differ widely, we scaled the values by log transformation (Eq. 1) and used the transformed values for BindingDB training.

$$IC_{50(scaled)} = -log_{10}(IC_{50}[M] + 10^{-10}) \quad (1)$$

LIGHTHOUSE architecture and training

The proposed overall model comprises two encoder networks (for chemicals and proteins) and one decoder network. MPNN is a message passing graph neural network that operates on compound molecular graphs[9]. In brief, MPNN conveys latent information among the atoms and edges. The message passing phase runs for t time steps and is defined in terms of message functions M_t and vertex update functions U_t. During this phase, hidden states h_v^t (128 dimensions in our model) at each node in the chemical graph are updated with the incoming messages m_v^{t+1} according to the following equations (Eqs. 2 and 3):

$$m_v^{t+1} = \sum_{w \in N(v)} M_t(h_v^t, h_w^t, e_{vw}) \quad (2)$$

$$h_v^{t+1} = U_t\left(h_v^t, m_v^{t+1}\right) \quad (3)$$

where e_{vw} represents edge feature between nodes v and w, $N(v)$ denotes the neighbor nodes of vertex v in graph G, and message functions M_t and update functions U_t are learned differentiable functions. After T (= 3) cycles of message passing and subsequent update, a readout function (average) is used to extract the embedding vectors at the graph level.

CNN is powerful for computer vision, but here we used a multilayer 1D CNN for protein sequence, as described previously[4]. In brief, the target amino acid is decomposed to each individual character and is encoded with an embedding layer and then fed into the CNN convolutions. We used three consecutive 1D convolutional layers with an increasing number of filters, with the second layer having double and the third layer having triple the number of filters in the first layer (32, 64, and 96 filters for the three layers). The convolution layers are followed by a global max-pooling layer. follows a global max-pooling layer. AAC is an 8,420-length vector in which each position corresponds to a sequence of three amino acids[11]. Transformer uses a self-attention–based transformer encoder[10] that operates on the substructure partition fingerprint of proteins. Algorithmically speaking, Transformer follows $O(n^4)$ in computation time and memory, where n is the input size[10]. This bottleneck prevented us from considering each amino acid as a token. We therefore used partition fingerprints to decompose amino acid sequence into protein substructures of moderate size and then fed each of the partitions into the model as a token[5].

As for the decoder, we exploited a previously described architecture[4]. In brief, encoder outputs are concatenated and entered into a three-layer feed-forward dense neural network (1024,1024, and 512 nodes), which finally outputs one value. We used Rectified Linear Unit (ReLU)[50], $g(x) = \max(0,x)$, as the activation function in the decoder network.

We defined our loss function with MSE (Eq. 4):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(P_i - Y_i)^2 \quad (4)$$

where P_i is the LIGHTHOUSE-predicted score for the ith compound-protein pair and Y_i is the true label in the corresponding training data, with a batch size of 128. We trained three architectures (MPNN_CNN, MPNN_AAC, MPNN_Transformer) separately for the STITCH and BindingDB training data with the Adam optimizer and a learning rate of 0.001. For evaluation metrics, we used MSE, concordance index, and Pearson correlation as well as AUROC. For every 10 epochs, we compared the current loss (in the validation data set) with that of 10 epochs ago; if the loss was not decreasing, we terminated the training for that model. As a result of this early termination, we trained MPNN_CNN for 40 epochs, MPNN_AAC for 70 epochs, and MPNN_Transformer for 100 epochs with regard to the confidence score (Extended Data Fig. 1b–g). As for the models for the interaction score, we trained MPNN_CNN for 70 epochs, MPNN_AAC for 100 epochs, and MPNN_Transformer for 70 epochs (Extended Data Fig. 2b–g), according to the same guidelines. After the training was completed, we finally evaluated the models with the test data sets, which were kept aside during the training and so had not previously been seen by the models.

Generation of virtual chemical libraries and prediction by LIGHTHOUSE

We prepared nearly 1 billion purchasable substances, which were downloaded from the ZINC database[23] as of 30 July 2020, for virtual PPAT inhibitor screening. For drug repurposing, we obtained approved drugs from the KEGG-DRUG database[51] as of 24 January 2021. For calculation of confidence and interaction scores, we fixed the proteins of interest (PPAT or ACE2) and changed the compounds iteratively, which yielded lists of predicted scores for all the compounds tested.

PPAT activity assay

Sf21 cells were cultured in Sf-900TM II SFM (Gibco, Cat# 10902-088) supplemented with 10 μM ferric ammonium citrate. They were transfected with a bacmid encoding human PPAT for 64 h, harvested, washed three times with phosphate-buffered saline, and lysed in a solution containing 150 mM NaCl, 25 mM Tris-HCl (pH 7.4), 0.5% Triton X-100, and 5 mM EDTA. The lysate was centrifuged at 10,000 × *g* for 6 min at 4°C, and the resulting supernatant (100 ng/ml) was incubated for 4 h at 37°C together with 5 mM glutamine (Gibco, Cat# 25030-081), 1 mM phosphoribosyl pyrophosphate (Sigma, Cat# P8296),

10 mM $MgCl_2$, 50 mM Tris-HCl (pH 7.4), and various concentrations of riboflavin 5′-monophosphate (Sigma, Cat# F2253-10 mg). Enzyme activity was assessed on the basis of glutamate production as measured with a glutamate assay kit (Abcam, Cat# 138883). The IC_{50} value was estimated from biological quadruplicates with a four-parameter logistic model[52] and with the use of JMP pro 15 software (version 15.1.0).

Assay of *E. coli* growth

Portions (20 μl) of *E. coli* strain JM109 (1×10^{10} colony-forming units (CFU)/ml) were cultured for various times in 2 ml of 2xYT liquid medium (BD Difco, Cat# 244020) containing various concentrations of pyridoxal 5′-phosphate (pH 7.0), after which OD_{600} was measured with a GENESYS 30 visible spectrophotometer (ThermoFisher Scientific, Cat# 840-277000). In addition, the JM109 strain was transformed with 1 μg of the pBlueScript II SK+ plasmid (Invitrogen), which harbors an ampicillin resistance gene as a selection marker, and was then spread on LB agar plates containing ampicillin (100 μg/ml) (Wako, Cat# 012-23303) with or without pyridoxal 5′-phosphate (3 mg/ml) and incubated overnight.

Virtual identification of statin targets and enrichment analyses

Three representative statins were fixed as chemical inputs, and all human protein-coding genes in the UniProt database were iteratively changed. The harmonic mean of the three confidence scores was calculated as an affinity score for statins, and the human protein-coding genes were sorted on the basis of this score. The resulting top 500 potential targets were then subjected to enrichment analyses with the use of the Metascape Web server[53].

SARS-CoV-2 assays

Vero-TMPRSS2 cells[54] were maintained in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum. The WK-521 strain of SARS-CoV-2 (EPI_ISL_408667) as well as the alpha (QK002, EPI_ISL_768526), beta (TY7-501, EPI_ISL_833366), gamma (TY8-612, EPI_ISL_1123289), and delta (TY11-927, EPI_ISL_2158617) variants were obtained from National Institute of Infectious Diseases in Japan. Stocks of these viruses were prepared by inoculation of Vero-TMPRSS2 cell cultures as described previously[54]. The MTT assay was performed to evaluate cell viability

after virus infection also as previously described[54]. In brief, serial twofold dilutions of ethoxzolamide in minimum essential medium (MEM) supplemented with 2% fetal bovine serum were added in duplicate to 96-well microplates. Vero-TMPRSS2 cells infected with wild-type or variant SARS-CoV-2 at 4 to 10 $TCID_{50}$ (median tissue culture infectious dose) were also added to the plates, which were then incubated at 37°C for 3 days. The viability of the cells was then determined with the MTT assay, and the culture supernatants were harvested for determination of the $TCID_{50}$ value as a measure of viral load. For indirect immunofluorescence analysis, cells infected with wild-type SARS-CoV-2 at a multiplicity of infection (MOI) of 0.0001 were cultured in the presence of various concentrations of ethoxzolamide for 64 h, fixed with 3.7% buffered formaldehyde, permeabilized with 0.05% Triton X-100, and incubated with antibodies to SARS-CoV-2 N protein (GeneTex, Cat# GTX635679). Immune complexes were detected with Alexa Fluor Plus 488–conjugated goat antibodies to rabbit immunoglobulin G (Invitrogen–ThermoFisher Scientific, Cat# A32731). Nuclei were stained with Hoechst 33342 (Invitrogen). Fluorescence images were captured with an IX73 fluorescence microscope (Olympus).

Methods references

50. Shimizu, H. & Nakayama, K.I. Artificial intelligence in oncology. *Cancer Sci.* **111**, 1452–1460 (2020).

51. Kanehisa, M., Furumichi, M., Sato, Y. Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551 (2021).

52. Pries, V. *et al*. Target identification and mechanism of action of picolinamide and benzamide chemotypes with antifungal properties. *Cell Chem. Biol*. **25**, 279–290 (2018).

53. Zhou, Y. *et al*. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, 1523 (2019).

54. Sasaki, M. *et al*. SARS-CoV-2 variants with mutations at the S1/S2 cleavage site are generated in vitro during propagation in TMPRSS2-deficient cells. *PLoS Pathog.* **17**, e1009233 (2021).

ACKNOWLEDGEMENTS

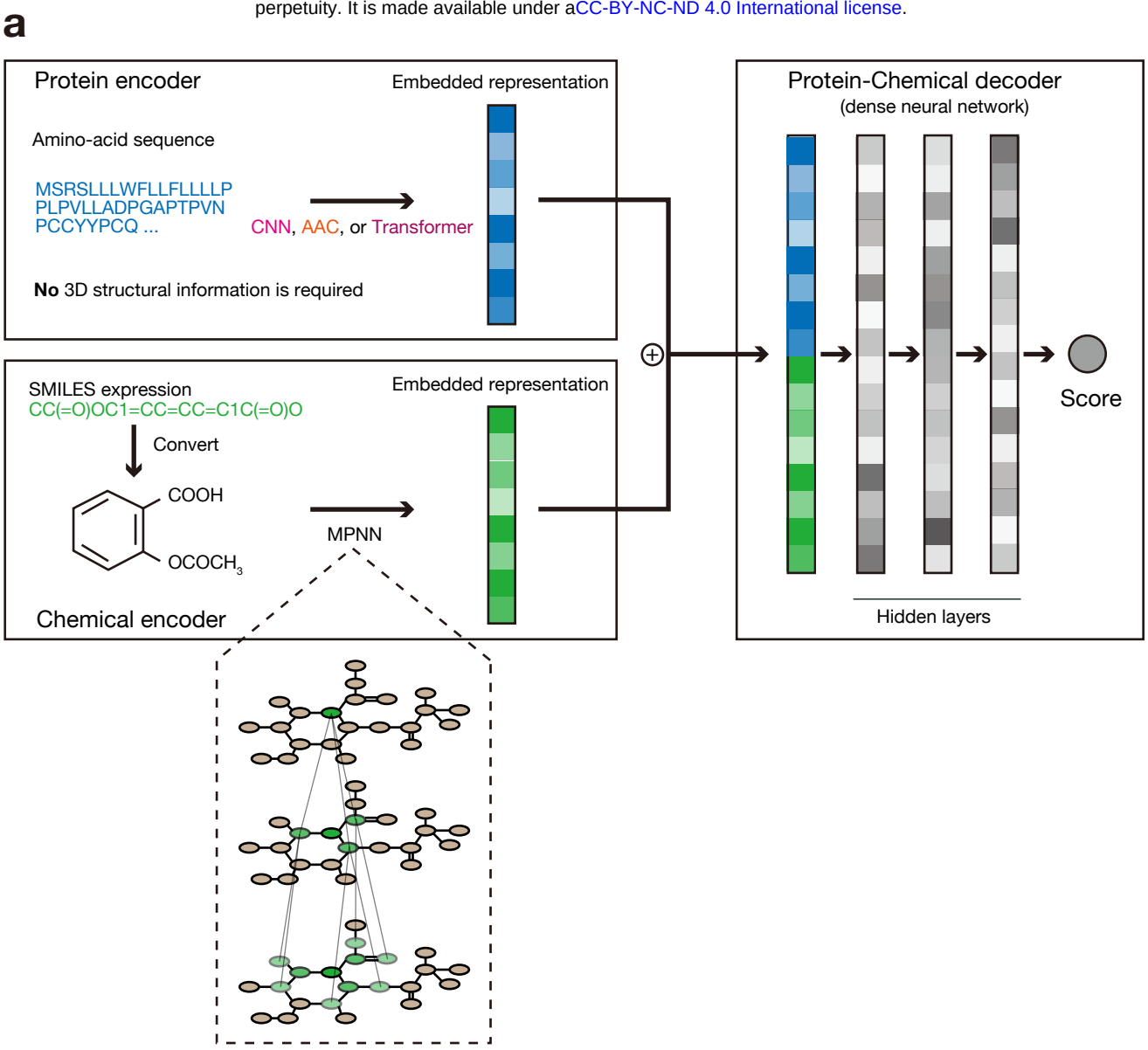
This work was supported by KAKENHI grants from the Japan Society for the Promotion of Science (JSPS) to H. Shimizu (21K17856), M.K. (21K15068), and K.I.N. (18H05215). H. Shimizu was also supported by postdoctoral fellowships from Takeda Science Foundation and Mochida Foundation. We thank T. Sawada and R. Wada for technical assistance, P.A. Silver and J.C. Way for critical reading of the manuscript, other laboratory members for discussion, and A. Ohta for help with preparation of the manuscript.

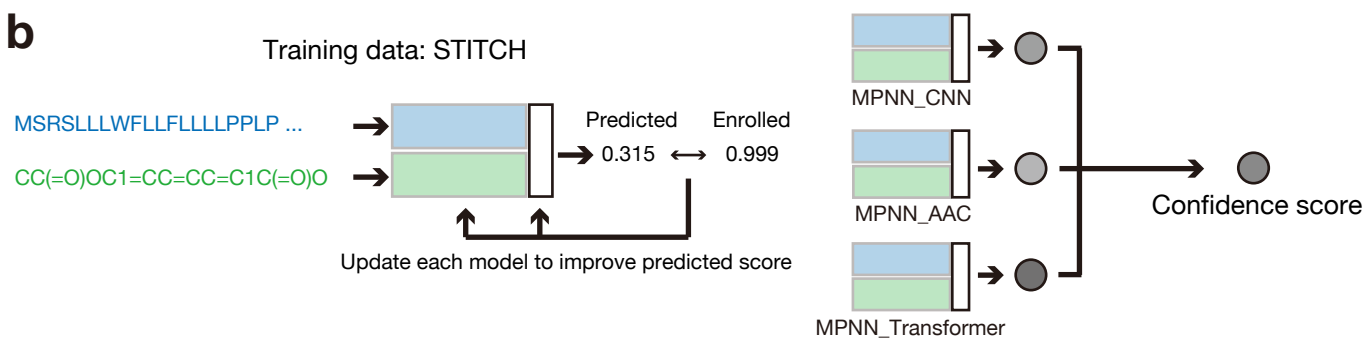
AUTHOR CONTRIBUTIONS

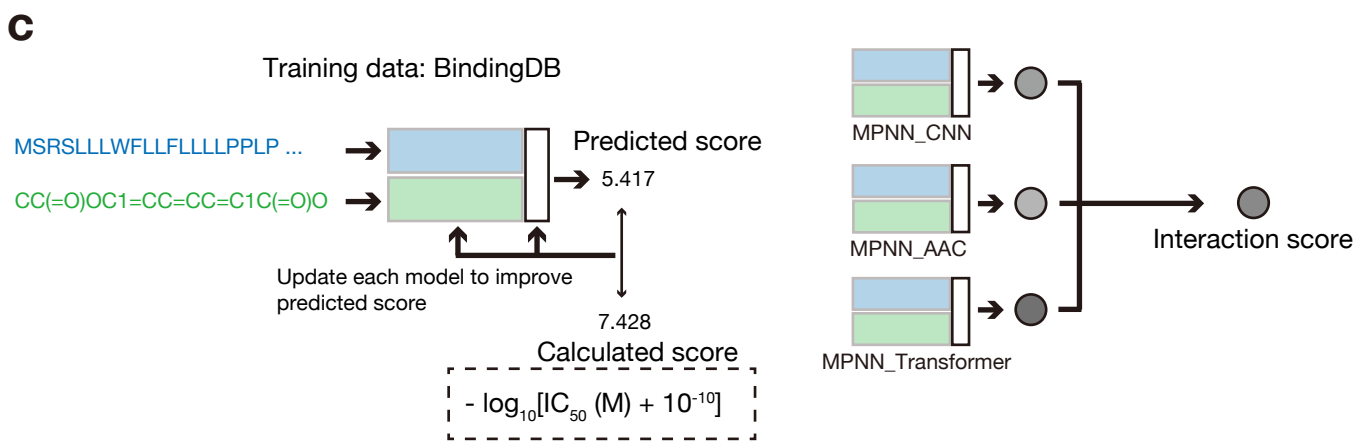
H. Shimizu conceived of and designed the study, developed LIGHTHOUSE, performed validation experiments, and wrote the original draft of the manuscript. M.K. conducted PPAT validation assays. Y.O., M.S., A.S., and H. Sawa performed COVID-19 infection analyses. M.M. contributed to discussion. K.I.N. supervised the study and edited the manuscript.

COMPETING INTERESTS

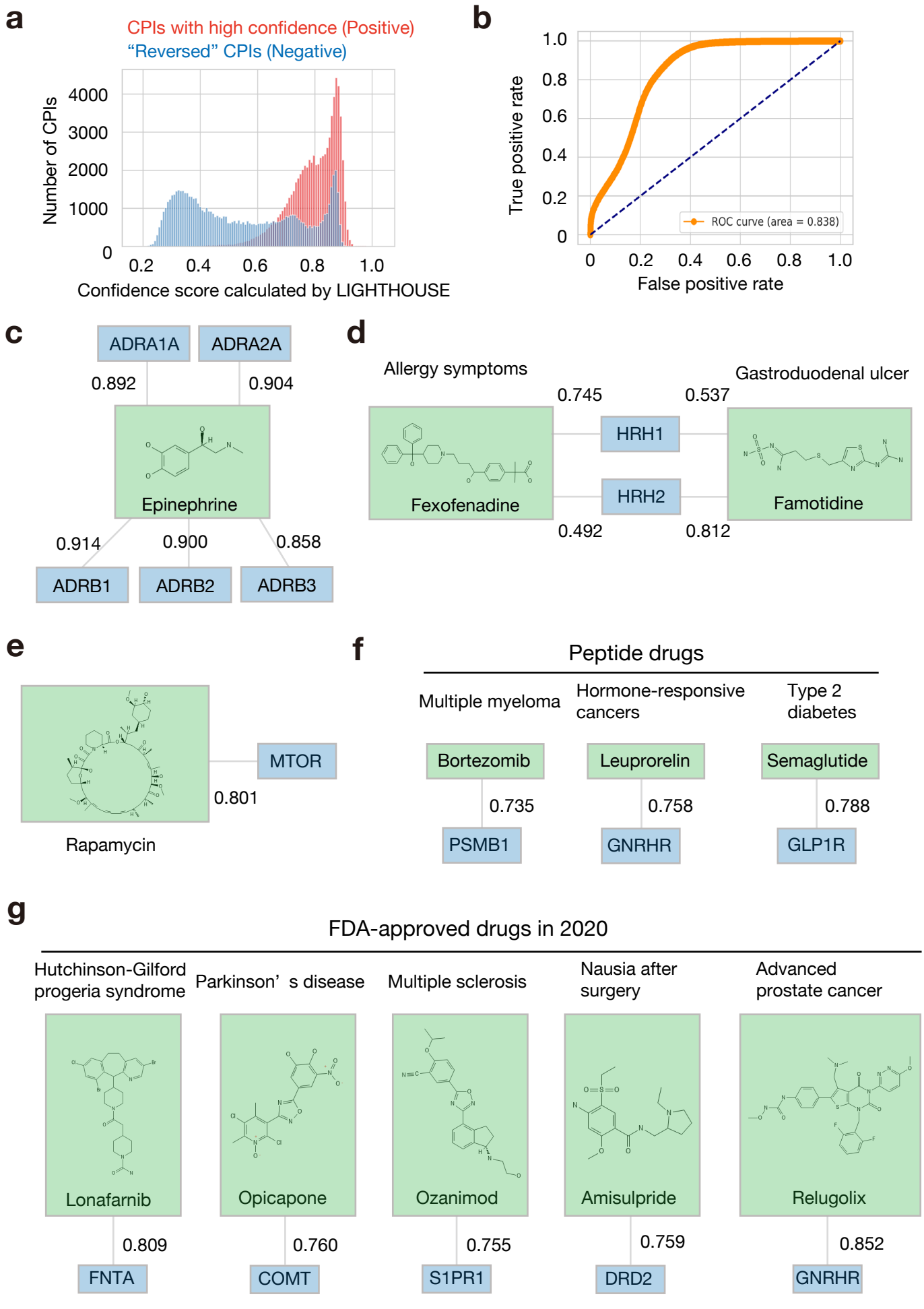
The authors declare no competing interests.



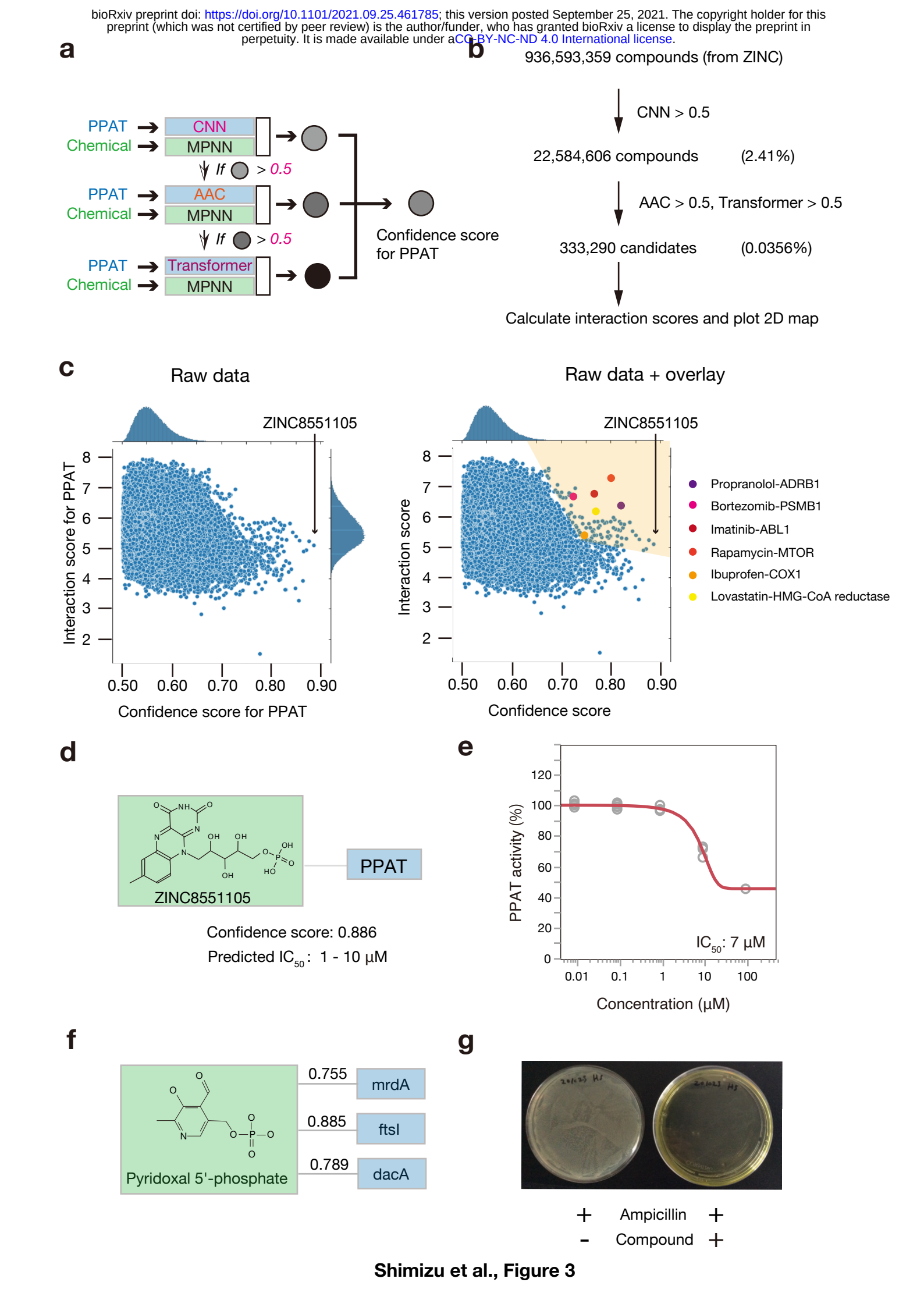




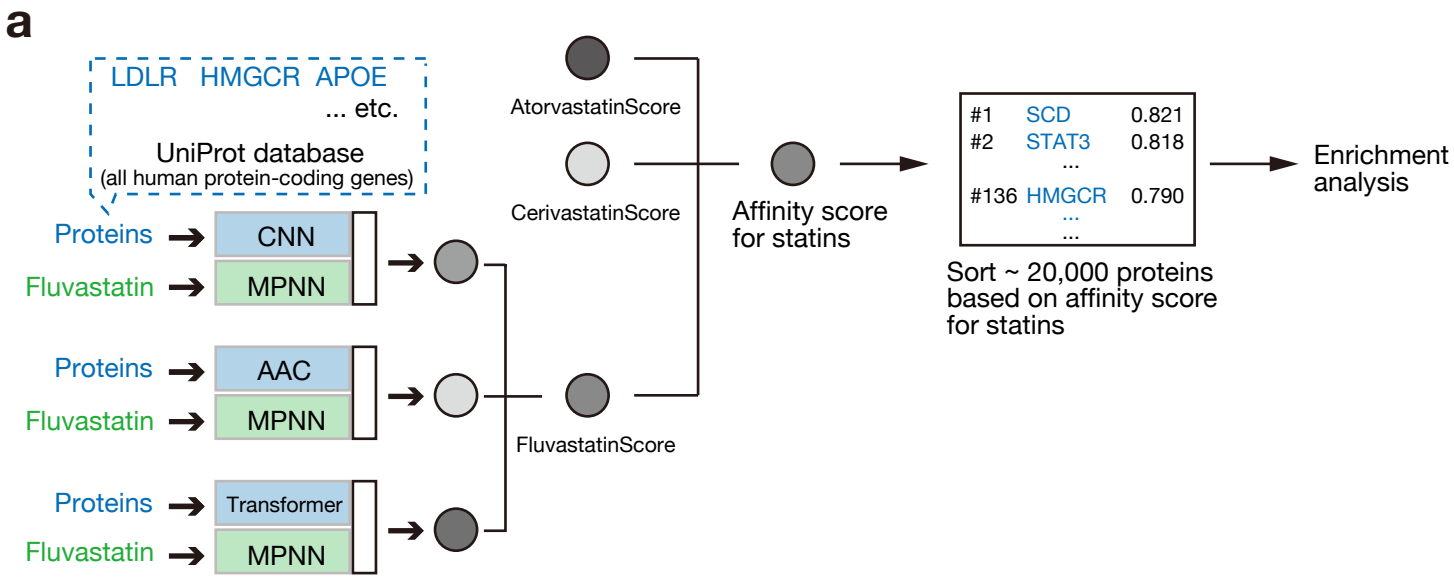
Shimizu et al., Figure 1

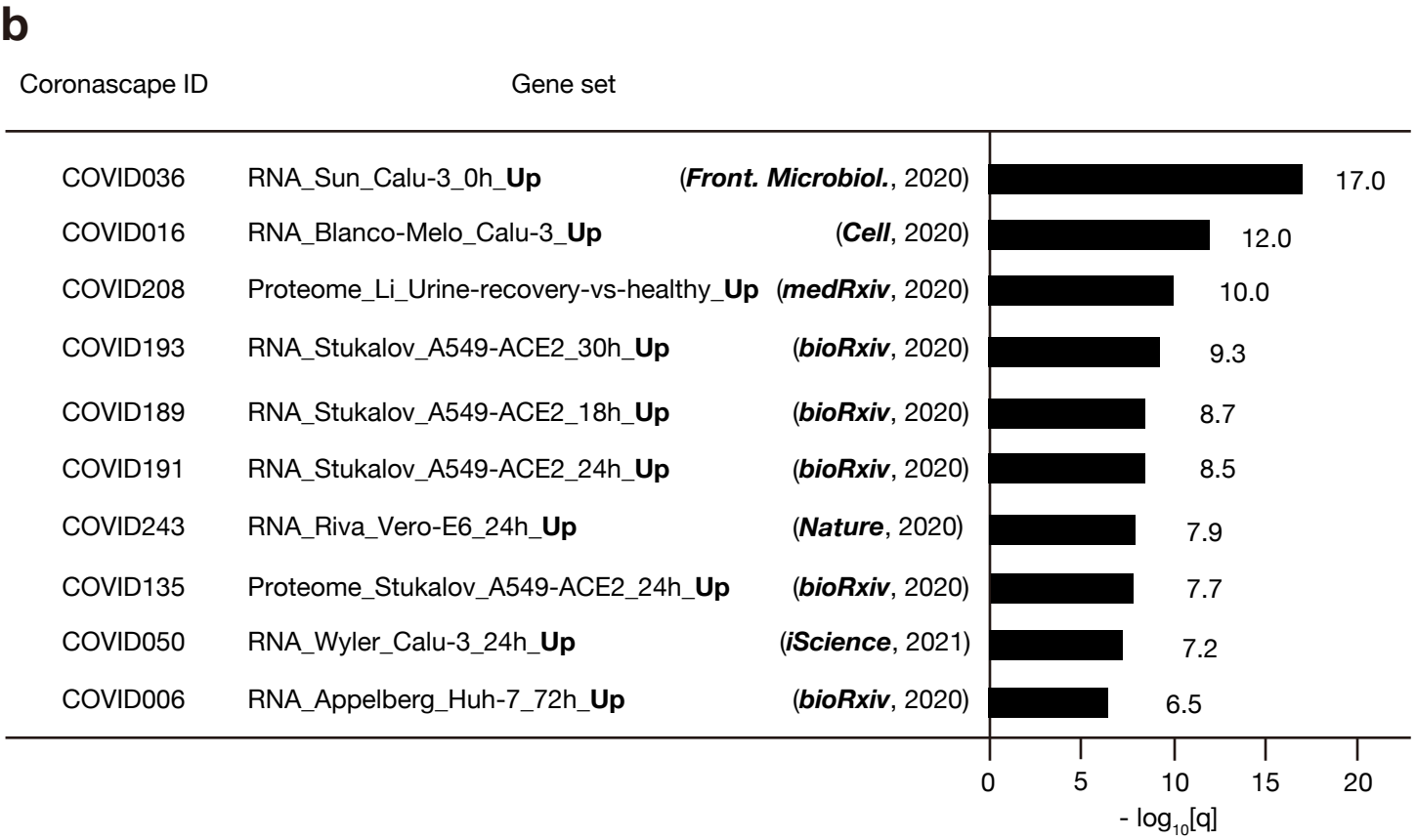


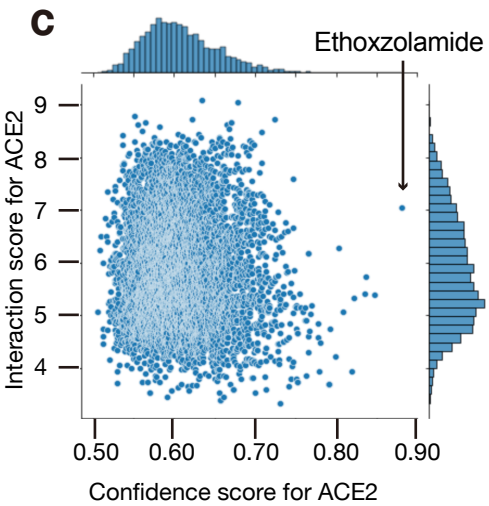
Shimizu et al., Figure 2

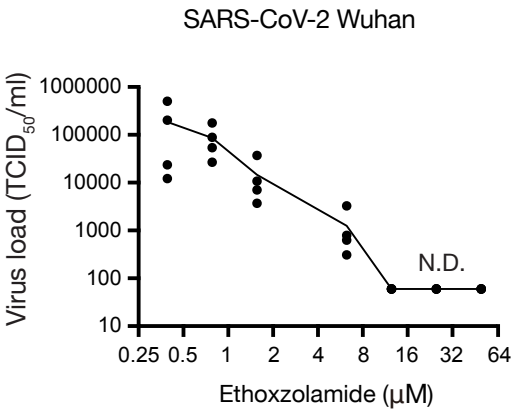


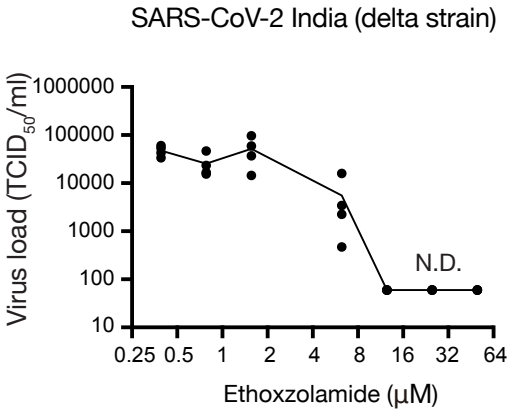
Shimizu et al., Figure 3











Shimizu et al., Figure 4