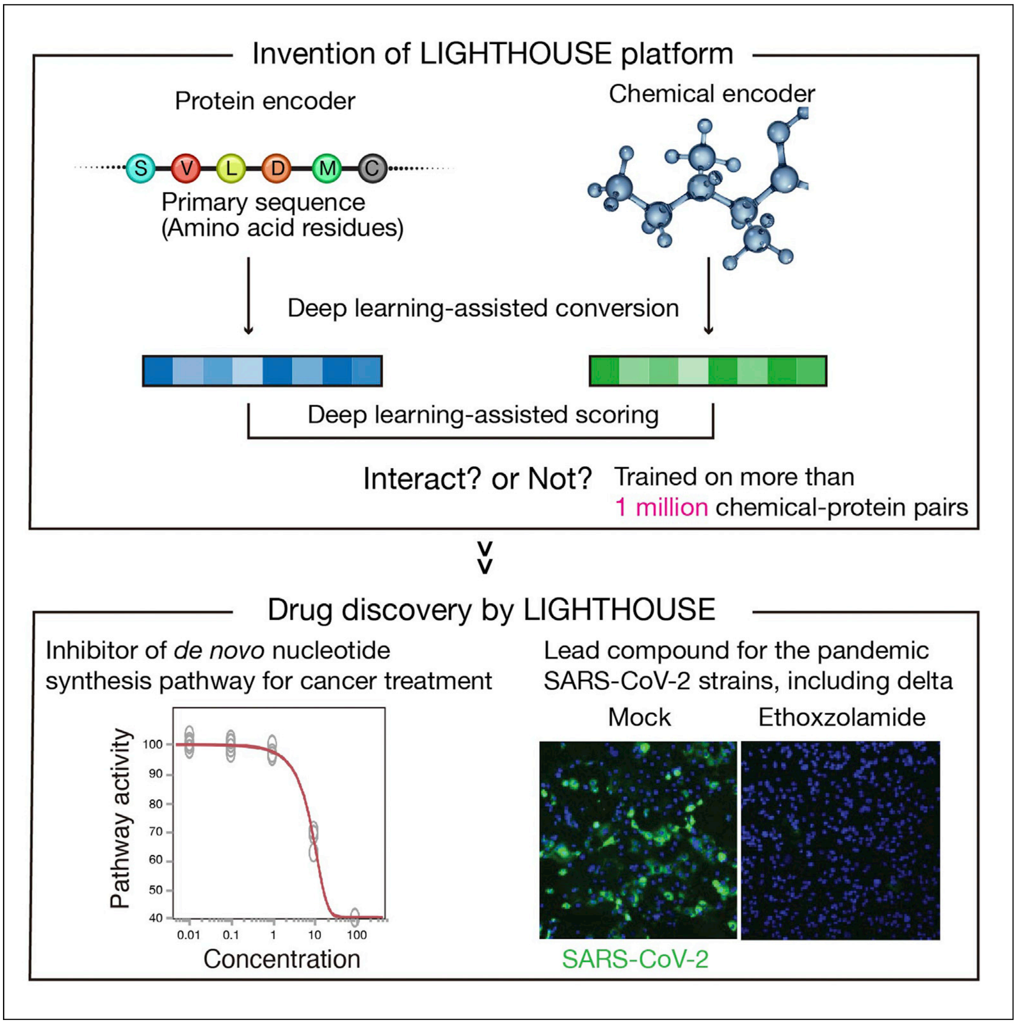
iScience

CellPress
OPEN ACCESS

Article

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


Invention of LIGHTHOUSE platform
Protein encoder
Chemical encoder
S
V
L
D
M
C
Primary sequence
(Amino acid residues)
Deep learning-assisted conversion
Deep learning-assisted scoring
Interact? or Not?
Trained on more than
1 million chemical-protein pairs
Drug discovery by LIGHTHOUSE
Inhibitor of *de novo* nucleotide
synthesis pathway for cancer treatment
Pathway activity
100
90
80
70
60
50
40
0.01
0.1
1
10
100
Concentration
Lead compound for the pandemic
SARS-CoV-2 strains, including delta
Mock
Ethoxzolamide
SARS-CoV-2

Hideyuki Shimizu, Manabu Kodama, Masaki Matsumoto, ..., Akihiko Sato, Hirofumi Sawa, Keiichi I. Nakayama

h_shimizu.dsc@tmd.ac.jp
(H.S.)
nakayak1@bioreg.kyushu-u.ac.jp (K.I.N.)

Highlights

LIGHTHOUSE discovers therapeutics solely on the basis of the primary sequence

The predictions of LIGHTHOUSE against multiple diseases were experimentally correct

LIGHTHOUSE facilitates optimization of lead compounds as well

Shimizu et al., iScience *25*, 105314
November 18, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.isci.2022.105314

iScience

CellPress
OPEN ACCESS

Article

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

Hideyuki Shimizu,[1,2,3,4,*] Manabu Kodama,[1] Masaki Matsumoto,[5] Yasuko Orba,[6] Michihito Sasaki,[6] Akihiko Sato,[6,7] Hirofumi Sawa,[6,8,9,10,11] and Keiichi I. Nakayama[1,12,*]

SUMMARY

One of the bottlenecks in the application of basic research findings to patients is the enormous cost, time, and effort required for high-throughput screening of potential drugs for given therapeutic targets. 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. 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 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. We envision that LIGHTHOUSE will help accelerate drug discovery and fill the gap between bench side and 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 progress slowly (Klein et al., 2017), 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 (Dobson, 2004), but HTS in most cases has been performed with only $\sim 10^{6}$ compounds. 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. More recently, with the advent of AlphaFold2 (Jumper et al., 2021), the ability to predict protein structures has been greatly improved, but it remains difficult to identify pockets of proteins that are potential drug targets and drug discovery without three-dimensional (3D) structural information therefore remains a challenge. Given that high-resolution 3D structural data are not available for most proteins to date and the high computational requirements of molecular docking simulations, the application of this approach has been limited.

Mathematical approaches have gained popularity in various fields, including sensing technologies (Greybush et al., 2019; Mohammadi Estakhri et al., 2019), clinical stratification (Shimizu and Nakayama, 2021), and other medical areas (Rajkomar et al., 2019). In particular, recent advances in artificial intelligence (AI) have demonstrated its potential in the pharmaceutical industry (Paul et al., 2021). Although many AI-based drug discovery methods have been proposed, they have had limited success in translational medicine. Whereas some studies have presented AI models with biological validation experiments (Stokes et al., 2020) or *de novo* molecular design (Zhavoronkov et al., 2019), many others have performed only computer-based validation without proof-of-concept biomedical experiments (Huang et al., 2021; Öztürk et al., 2018; Tsubaki et al., 2019). In addition, most platforms to date have been trained with small datasets, such as Directory of Useful Decoys Enhanced (DUD-E), that have known biases (Chen et al., 2019) 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, might be expected to be more accurate and appropriate for AI-based drug discovery (Hansen and Salamon, 1990).

[1] Department of Molecular and Cellular Biology, Medical Institute of Bioregulation, Kyushu University, Fukuoka 812-8582, Japan

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

[3] Wyss Institute for Biologically Inspired Engineering, Harvard Medical School, Boston, MA 02115, USA

[4] Department of AI Systems Medicine, M&D Data Science Center, Tokyo Medical and Dental University, Tokyo 113-8510, Japan

[5] Department of Omics and Systems Biology, Niigata University Graduate School of Medical and Dental Sciences, Niigata 951-8510, Japan

[6] Division of Molecular Pathobiology, International Institute for Zoonosis Control, Hokkaido University, Sapporo 060-8638, Japan

[7] Drug Discovery and Disease Research Laboratory, Shionogi & Co. Ltd., Osaka 561-0825, Japan

[8] International Collaboration Unit, International Institute for Zoonosis Control, Hokkaido University, Sapporo 060-8638, Japan

[9] One Health Research Center, Hokkaido University, Sapporo 060-8638, Japan

[10] Global Virus Network, Baltimore, MD 21201, USA

[11] Hokkaido University, Institute for Vaccine Research and Development (HU-IVReD)

[12] Lead contact

*Correspondence:
h_shimizu.dsc@tmd.ac.jp
(H.S.),
nakayak1@bioreg.kyushu-u.ac.jp (K.I.N.)

https://doi.org/10.1016/j.isci.2022.105314

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.

Nevertheless, AI-driven drug discovery continues to gain momentum and achieve critical milestones, especially in industry. The first AI-designed drug candidates to enter clinical trials were reported by Exscientia in early 2020. All three molecules (DSP-1181, EXS21546, and DSP-0038) are in phase 1 trials and were discovered with Exscientia's AI platform (Jayatunga et al., 2022). DSP-1181 is a full agonist of the 5-HT1a serotonin receptor that was discovered in a collaboration between Exscientia and Sumitomo Dainippon Pharma, and EXS21546 is an A2a adenosine receptor antagonist discovered in a collaboration between Exscientia and Evotec. Another example is ISM001-055, a small-molecule inhibitor aimed at idiopathic pulmonary fibrosis. The target and the drug were identified by using AI (Kirkpatrick, 2022). Several other AI-based small-molecule drug candidates are also now in clinical trials (Jayatunga et al., 2022). However, technological details have not been disclosed, and the AI systems are not available to other researchers.

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 promote 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.

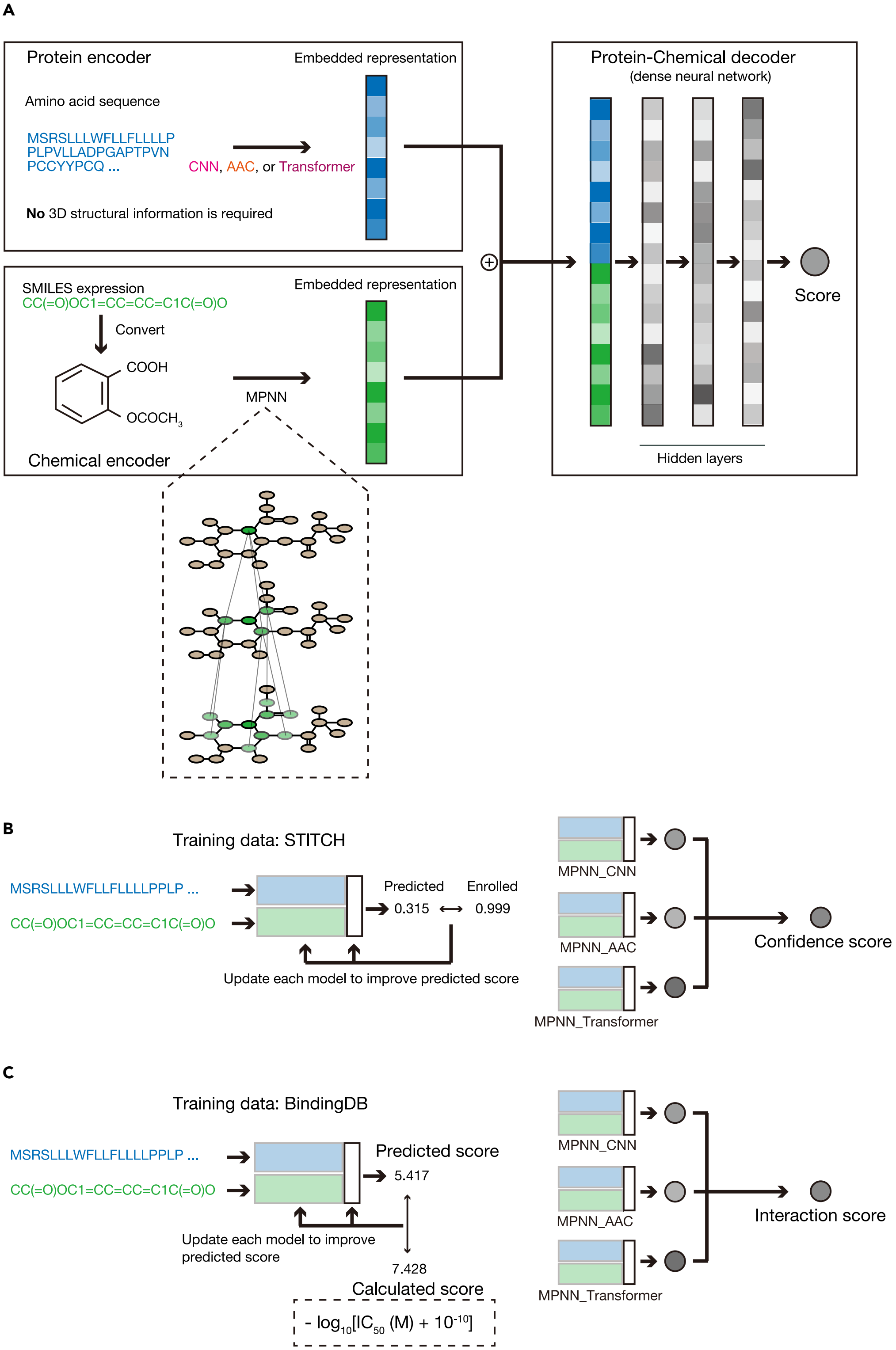
RESULTS

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 (Gilmer et al., 2017) 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) (Öztürk et al., 2018), Transformer (Vaswani et al., 2017), and AAC (amino acid composition up to 3-mers) (Reczko and Bohr, 1994). 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 (Figure 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 (Table S1), which is one of the largest CPI databases (Szklarczyk et al., 2016) 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 randomly divided the overall data into training (80%), validation (10%), and test (10%) datasets (Figure S1A).

We fed the network with protein primary structures and chemicals and trained it to output the scores from the STITCH training dataset (Figure 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 (Figures S1B–S1G). These findings indicated that our AI models learned the approximation of the hidden 1D relation underlying the compound-protein pairs without overfitting the training data. We examined the performance of the models with the test dataset at the end of the training and (epoch-wise) validation phases, and we discovered that the AUROC for all three models was >0.80 (Table S2). These scores are



Shimizu et al., Figure 1

Figure 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 and 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 $\sim$100 μM, whereas an interaction score of 9 means that, if the compound has inhibitory activity, the IC_{50} would be $\sim$1 nM. Note that module 2 only works if the compound and protein interact, so this module is auxiliary to module 1.

See also Figures S1, S2 and Tables S1, S2, S3.

equivalent to or better than those of cutting-edge 3D docking simulations (Hsin et al., 2016; Moussa et al., 2021; Wang et al., 2020). 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 (Figure 1B).

We also trained the models to predict scores based on IC_{50} values. For this purpose, we used data from BindingDB (Gilson et al., 2016), which collects a variety of experimental findings, and we divided the data into training (80%), validation (10%), and test (10%) datasets (Figure S2A). The same architecture was adopted to train the AI models to predict scaled IC_{50} values (Figure 1C), yielding an interaction score, and we confirmed that the models adequately learned how to predict IC_{50} from amino acid sequence–chemical pairs (Figures S2B–S2G). Finally, we assessed the performance of the models with undisclosed test data, finding that they performed well in predicting IC_{50} (Table S3).

LIGHTHOUSE architecture outperforms state-of-the-art methods

We next compared the performance of LIGHTHOUSE with that of similar existing methods. To ensure a fair comparison, we used DUD-E data as were used in previous studies (Tsubaki et al., 2019; Wallach et al., 2015). In brief, we randomly split DUD-E data (102 target proteins) into training (72 proteins) and test (30 proteins) data and then trained the LIGHTHOUSE architecture with the DUD-E training data for classification of compounds as active or decoy with regard to the protein in question. After this training, we examined LIGHTHOUSE performance with the DUD-E test data (Figure S3A). Of note, we used only amino acid sequences of the proteins for training and evaluation of LIGHTHOUSE, even though DUD-E provides structural data (as PDB files) for proteins. In addition, we used the balanced dataset of DUD-E—the training samples comprise 22,886 active (positive) and 22,886 decoy (negative) samples—as in a previous study (Tsubaki et al., 2019). LIGHTHOUSE yielded an AUROC for the DUD-E test data of 0.956 (Figure S3B), which was higher than the values produced by state-of-the-art methods including 3D-CNN (Ragoza et al., 2017), AtomNet (Wallach et al., 2015), and a graph-based deep learning method proposed by Tsubaki et al. (2019) (Figure S3C).

For further comparison of LIGHTHOUSE with the second best method, we downloaded CPI data for human and *Caenorhabditis elegans* from the GitHub repository of Tsubaki et al. (2019). Both of these datasets were generated previously (Liu et al., 2015). Similar to Tsubaki et al., we retrained the LIGHTHOUSE architecture with these training data, and we found that LIGHTHOUSE outperformed this other method on the basis of both AUROC and F1 metrics (Table S4). These bodies of evidence thus show that LIGHTHOUSE is one of the best architectures for drug discovery available to date.

In silico verification of LIGHTHOUSE

We next evaluated the performance of LIGHTHOUSE in terms of its ability to predict known CPIs. We generated two datasets for this purpose: a "Positive" dataset consisting of reliable CPIs (STITCH

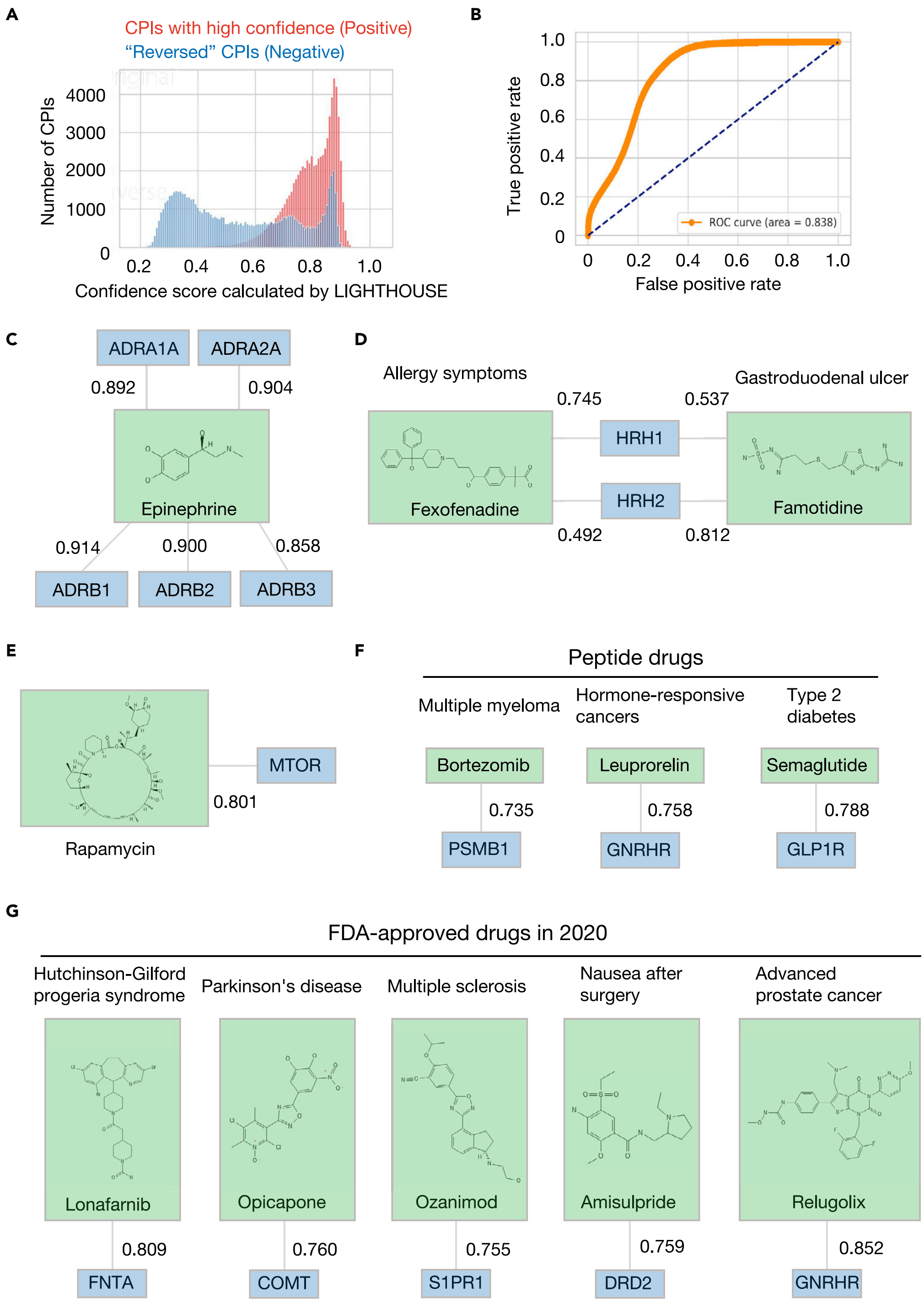


Figure 2. In silico verification of LIGHTHOUSE

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

Figure 2. *Continued*

(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. These peptide drugs were converted to numerical vectors with the use of SMILES expression and MPNN.

(G) Application of LIGHTHOUSE to five drugs approved by the FDA in 2020 that were not included in the training dataset (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.

See also Figures S5, S8 and Table S5.

confidence score of >0.9), and a ''Negative'' dataset in which the amino acid sequences of the ''Positive'' dataset 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 datasets revealed that those for the ''Positive'' dataset were heavily skewed toward 1 (Figure 2A). Receiver operating characteristic (ROC) curve analysis showed that the two datasets could be distinguished on the basis of their LIGHTHOUSE confidence scores (Figure 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'' dataset. We calculated confidence and interaction scores for ATP as well as for the tyrosine kinase inhibitor drugs sorafenib and sunitinib as true positive examples. ATP showed a skewed distribution with respect to the confidence score, indicating that it is a false positive compound (Figure S4A). In contrast, the distribution of confidence scores for both sorafenib and sunitinib (Figures S4B and S4C) was completely different from that for ATP. 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.

To demonstrate further the predictive power of LIGHTHOUSE, we searched all 10 known targets of sorafenib registered in the DrugBank database (Wishart et al., 2018). We found that 9 out of the 10 known targets were located in the promising compartment (confidence score of >0.7 and interaction score of >7.0) (Figure S5A), which constitutes statistically significant enrichment ($p = 2.4 \times 10^{-14}$, Fisher exact test). Furthermore, 8 of the 10 known sorafenib targets were among the top 25 candidate proteins (confidence score of >0.75 and interaction score of >7.5) (Figure S5B). In addition to these 8 known sorafenib targets, the top 25 candidate proteins included an additional 2 kinases that may be unknown targets of this multikinase inhibitor. Thus, 40% of the top candidates (10 out of 25) were found to be experimentally verified or likely true positives.

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 (Figure 2C). Histamine receptors are classified into four subtypes (Seifert et al., 2013), with HRH1 and HRH2 being targets of anti-allergy 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 (Figure 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) (Figure 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) (Figure 2F), yielding high confidence scores for these drugs and their known targets: proteasome subunit PSMB1 (Berkers et al., 2005), gonadotropin-releasing hormone receptor (GNRHR) (Borroni et al., 2000), and glucagon-like peptide–1 (GLP-1) receptor (GLP1R) (Knudsen and Lau, 2019), respectively. Given the rapidly growing demand for peptide drugs (Muttenthaler et al., 2021), 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 US 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 (Figure 2G), indicating the expandability of LIGHTHOUSE to a much larger exploration space than that encompassed by STITCH.

Furthermore, we also evaluated IC_{50} data from BindingDB, which are derived from actual bioassays. We found that the predicted value and observed value correlate well for the BindingDB test dataset (Figure S6). 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 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 (Kodama et al., 2020). 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 $\sim 10^9$ commercially available compounds in the ZINC database (Sterling and Irwin, 2015) (Figure S7). To reduce the calculation time, we adopted a step-by-step application of LIGHTHOUSE (Figure 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 (Figure 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 (Figure 3C, left). We hypothesized that we would achieve better prediction with the use of both confidence and interaction scores, and the model indeed improved by ~17.9% as determined on the basis of the AUROC (Figures 2A, 2B, and S8).

The best candidates would be expected to have high confidence and interaction scores, appearing in the upper right corner of the 2D plot. Indeed, this criterion was met by several well-known drug-target combinations (Figure 3C, right). To thoroughly investigate known drug-protein interactions, we downloaded the "Model List of Essential Medicines" published by WHO (World Health Organization, 2021) and then searched the DrugBank database (Wishart et al., 2018) for each essential drug and excluded those without known target proteins. If multiple known target proteins were registered, the protein at the top of the target list was used as a representative example, and its amino acid sequence was obtained from UniProt, which is linked to DrugBank. Of the 112 drug-protein combinations that we calculated by LIGHTHOUSE, most (79 combinations, 70.5%) (Table S5) were located in the upper right compartment depicted in Figure 3C (right).

Among the >333,000 final compounds in the PPAT analysis, the top candidate PPAT inhibitor with the highest confidence score was ZINC8551105 (riboflavin 5′-monophosphate), with a predicted IC_{50} of 1–10 μM (Figure 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 7.5 μM (Figure 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. It is also of note that we tested only this compound, so other top candidates might also inhibit PPAT activity.

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 (Tooke et al., 2019) 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 (*ftsI*), and PBP5 (*dacA*), all of which are essential for cell wall synthesis in *Escherichia coli* (Macheboeuf et al., 2006) (Figure 4A). This compound indeed suppressed the growth of wild-type *E. coli* strain JM109 in a concentration-dependent manner (Figure 4B). Importantly, pyridoxal 5′-phosphate also markedly inhibited the growth of an ampicillin-resistant *E. coli* transformant that produces β-lactamase

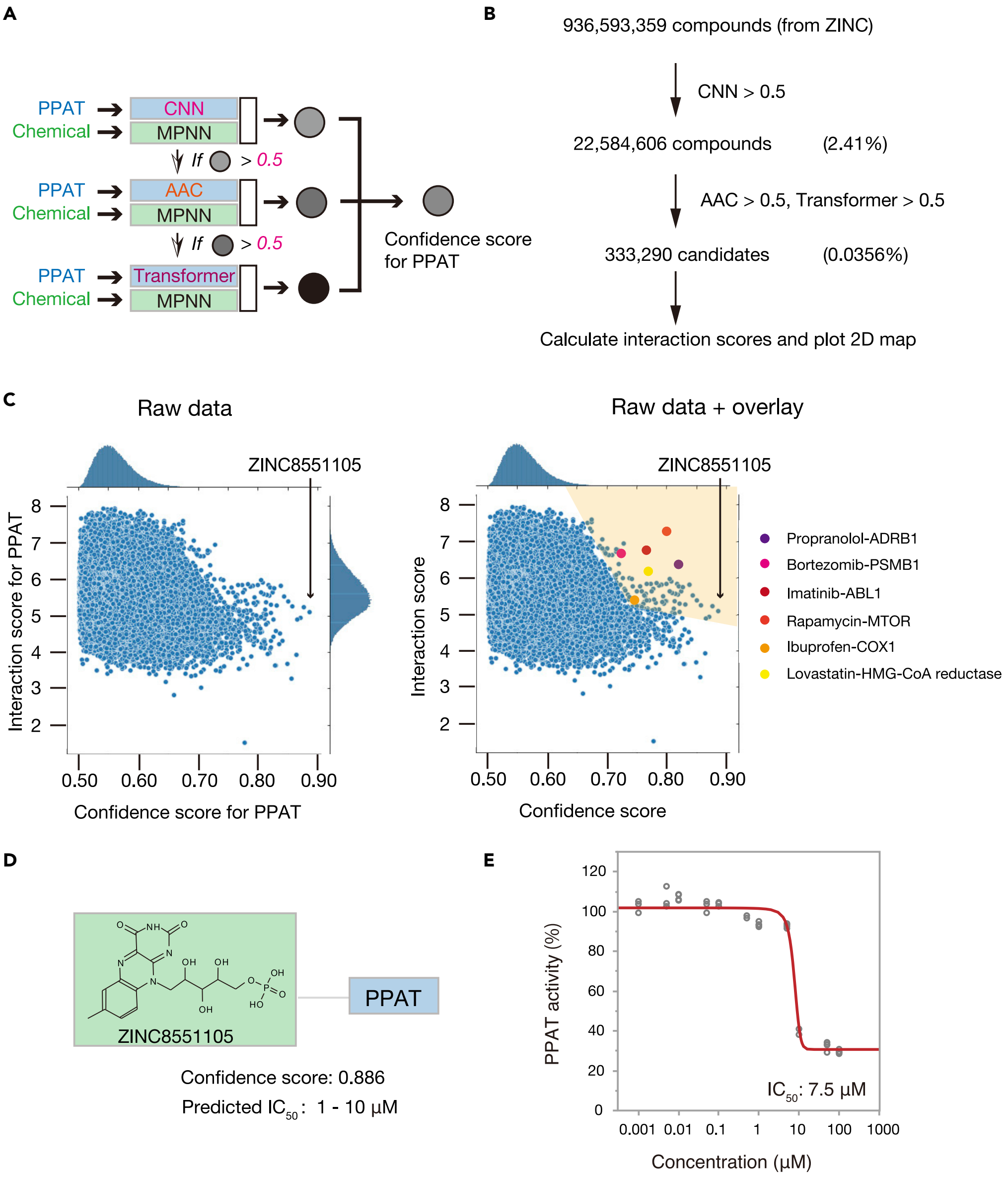


Figure 3. Discovery of lead compounds for treatment of cancer

(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 (1, 5, 10, 50, 100, 500 nM, 1, 5, 10, 50, and 100 µM) of riboflavin 5′-monophosphate, with the determined IC_{50} value being within the range predicted by LIGHTHOUSE. Data are shown for four biological replicates.

See also Figures S7 and S8.

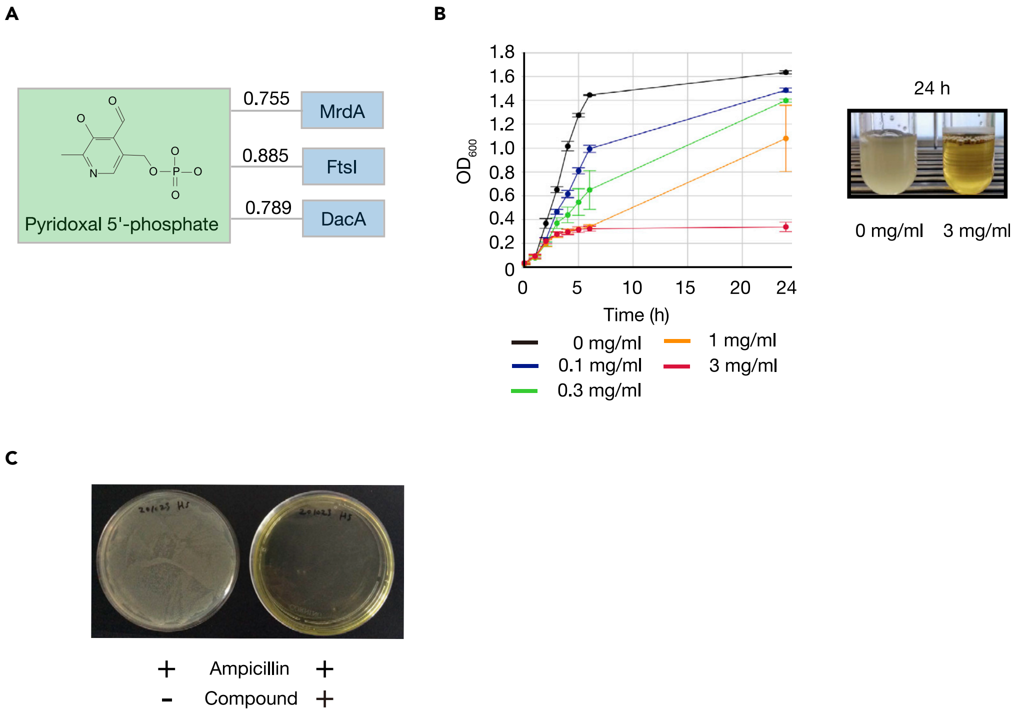


Figure 4. LIGHTHOUSE is applicable to prokaryotic proteins

(A) LIGHTHOUSE predicted that pyridoxal 5′-phosphate would associate with several penicillin binding proteins (PBPs) of *E. coli* (strain K12). MrdA and FtsI are peptidoglycan D,D-transpeptidases, whereas DacA is a D-alanyl-D-alanine carboxypeptidase.

(B) *Escherichia coli* strain JM109 was cultured in 2xYT medium supplemented with various concentrations of pyridoxal 5′-phosphate, and optical density at 600 nm (OD_{600}) of the culture was monitored. Data are means ± SD for three independent experiments.

(C) 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. The pH of pyridoxal 5′-phosphate was adjusted to 7.0 in order to avoid potential nonspecific toxicity.

(Figure 4C). 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 (Zheng et al., 2018). 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 (Deacon, 2020). 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 (Li et al., 2016) (Figure 5A). Furthermore, LIGHTHOUSE predicted that removal of the phosphate group would reduce the inhibitory potency of riboflavin 5′-monophosphate for PPAT (Figure 3E), and this prediction was confirmed correct by the finding that the IC_{50} value was increased from 7.5 to 49.9 μM (Figure 5B). These data suggested the possibility that LIGHTHOUSE is capable of predicting activity cliffs.

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 (Jabbour et al., 2008), and LIGHTHOUSE accurately predicted the effect of this mutation (Figure 5C). LIGHTHOUSE is able to provide such insight from only wild-type amino acid sequences, given the lack of variant information in the original

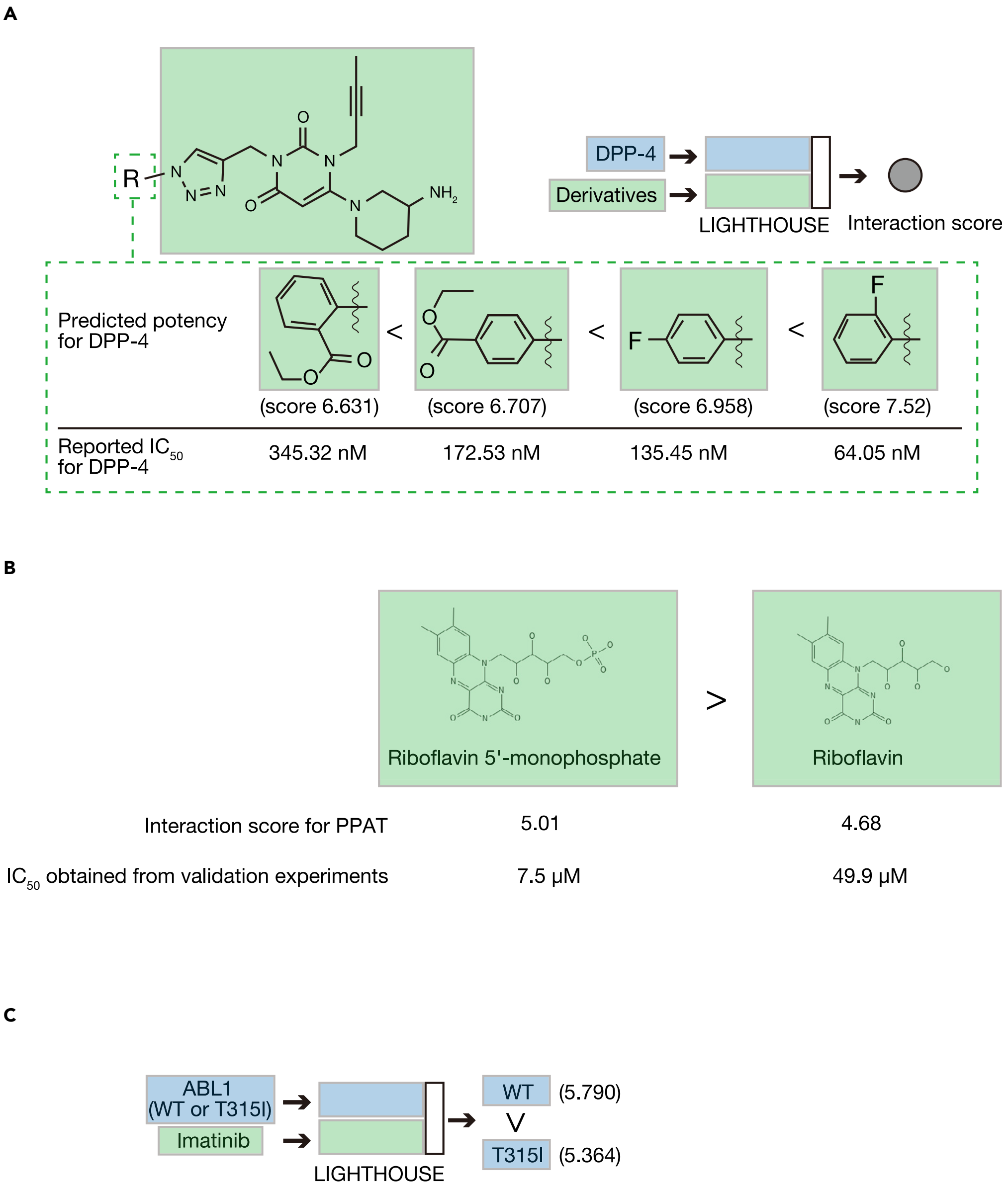


Figure 5. LIGHTHOUSE directs optimization of lead compounds

(A) Prediction of the potency of DPP-4 inhibitor derivatives by LIGHTHOUSE. The predicted interaction scores are compared with the reported IC_{50} values.

(B) LIGHTHOUSE accurately predicts that riboflavin is a less potent PPAT inhibitor than is riboflavin 5′-monophosphate.

(C) The interaction scores calculated by LIGHTHOUSE for both wild-type (WT) and T315I mutant forms of ABL1 accurately predict that the point mutation reduces the effectiveness of the leukemia drug imatinib.

training dataset. 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 might 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

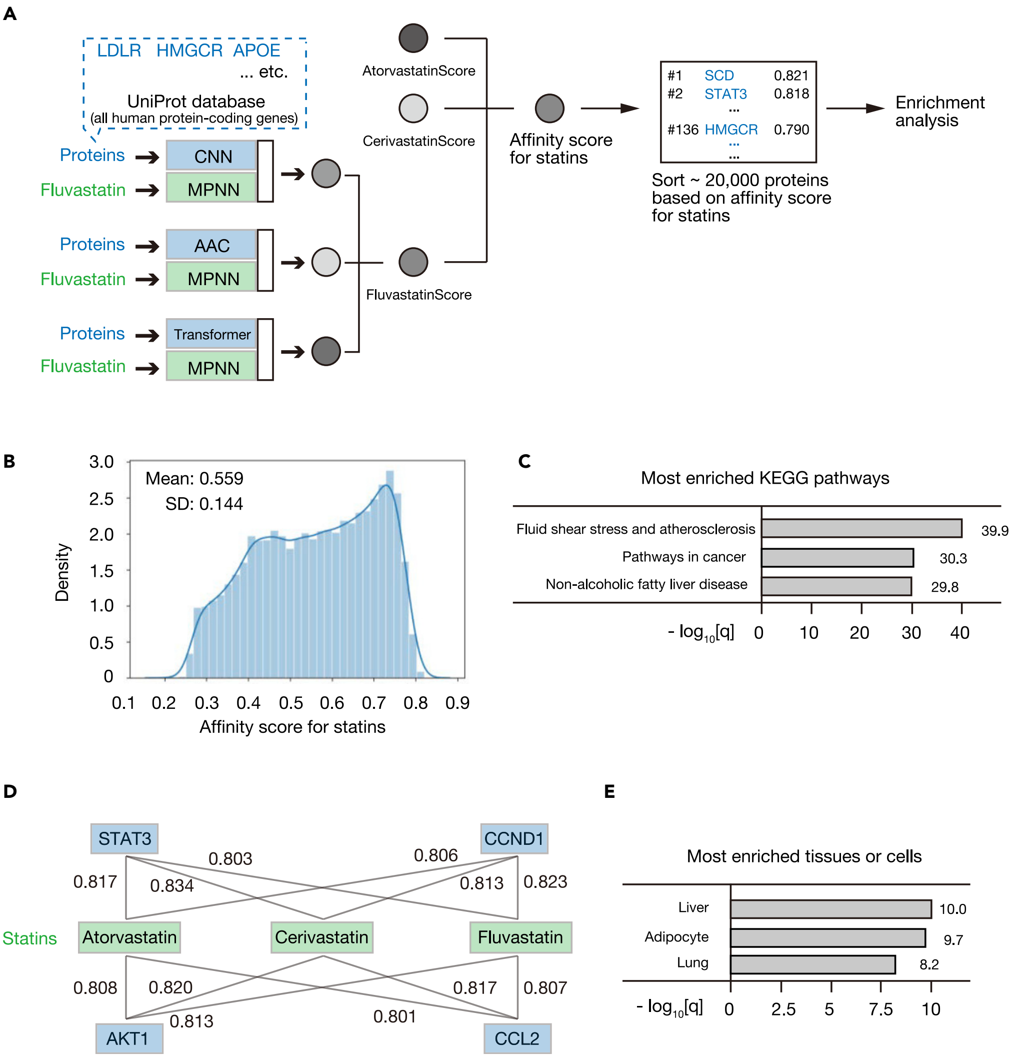


Figure 6. LIGHTHOUSE uncovers potential target proteins for given drugs

(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 (Fluvastatin Score, Atorvastatin Score, and Cerivastatin Score) 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) Distribution of the harmonic mean of the Atorvastatin Score, Cerivastatin Score, and Fluvastatin Score (affinity score for statins).

(C) KEGG pathway enrichment analysis for the top 500 potential statin targets identified by LIGHTHOUSE. Minus $\log_{10}$-transformed q values are shown.

(D) Confidence scores for representative predictions by LIGHTHOUSE of the association of statins with cancer-related proteins.

(E) Enrichment analysis for expression sites of the top 500 potential statin targets. Minus $\log_{10}$-transformed q values are shown. See also Table S6.

cancer, although the target molecules for these effects have remained unclear (Mei et al., 2017). We therefore applied LIGHTHOUSE to three representative statins (atorvastatin, cerivastatin, and fluvastatin) and computed confidence scores for all human protein-coding genes (Figures 6A and 6B, Table S6). 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

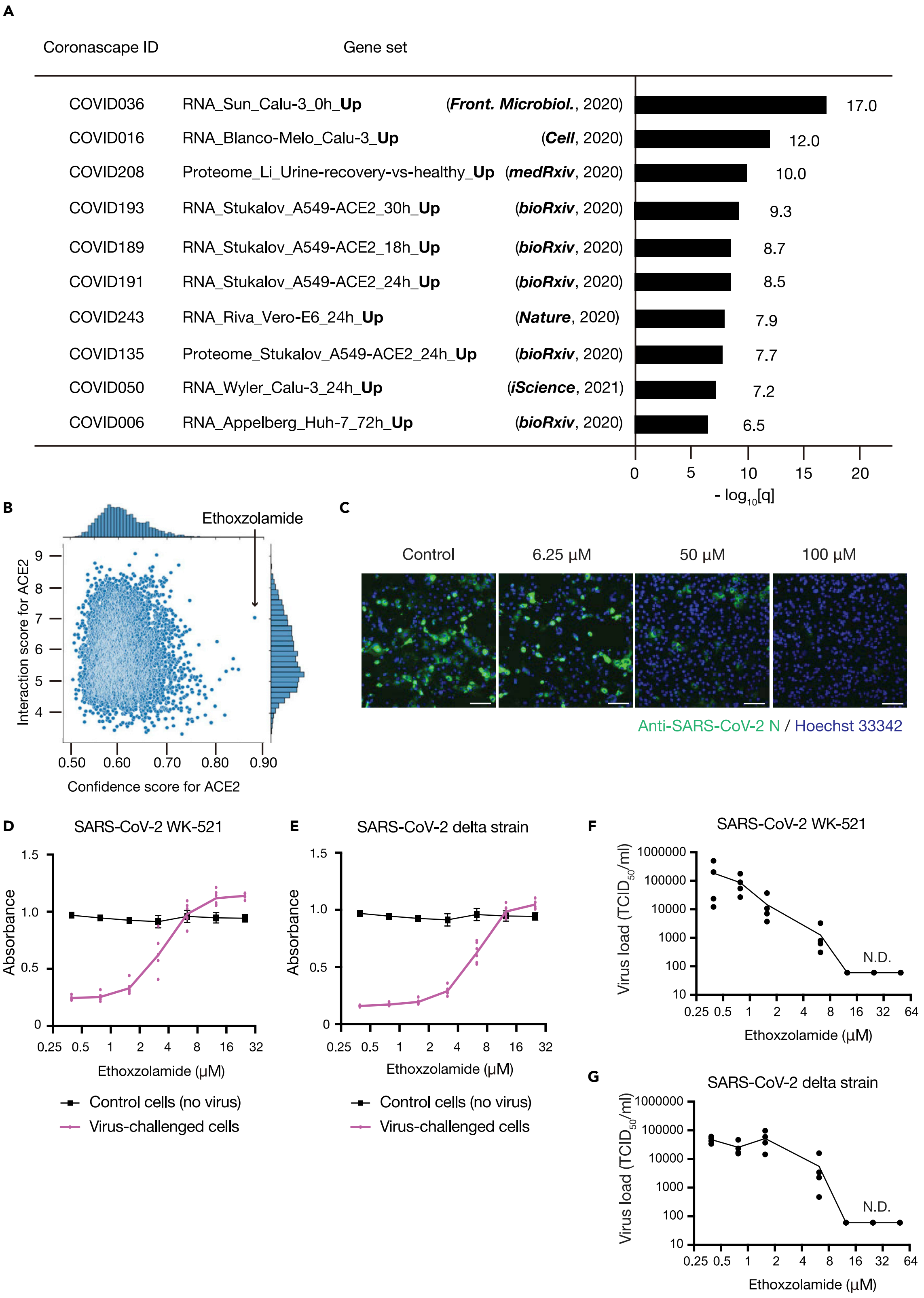

A
Coronascape ID
Gene set
COVID036
RNA_Sun_Calu-3_0h_Up
(Front. Microbiol., 2020)
17.0
COVID016
RNA_Blanco-Melo_Calu-3_Up
(Cell, 2020)
12.0
COVID208
Proteome_Li_Urine-recovery-vs-healthy_Up
(medRxiv, 2020)
10.0
COVID193
RNA_Stukalov_A549-ACE2_30h_Up
(bioRxiv, 2020)
9.3
COVID189
RNA_Stukalov_A549-ACE2_18h_Up
(bioRxiv, 2020)
8.7
COVID191
RNA_Stukalov_A549-ACE2_24h_Up
(bioRxiv, 2020)
8.5
COVID243
RNA_Riva_Vero-E6_24h_Up
(Nature, 2020)
7.9
COVID135
Proteome_Stukalov_A549-ACE2_24h_Up
(bioRxiv, 2020)
7.7
COVID050
RNA_Wyler_Calu-3_24h_Up
(iScience, 2021)
7.2
COVID006
RNA_Appelberg_Huh-7_72h_Up
(bioRxiv, 2020)
6.5
0
5
10
15
20
- log$_{10}$[q]
B
Ethoxzolamide
Interaction score for ACE2
9
8
7
6
5
4
0.50
0.60
0.70
0.80
0.90
Confidence score for ACE2
C
Control
6.25 µM
50 µM
100 µM
Anti-SARS-CoV-2 N / Hoechst 33342
D
SARS-CoV-2 WK-521
Absorbance
1.5
1.0
0.5
0
0.25
0.5
1
2
4
8
16
32
Ethoxzolamide (µM)
Control cells (no virus)
Virus-challenged cells
E
SARS-CoV-2 delta strain
Absorbance
1.5
1.0
0.5
0
0.25
0.5
1
2
4
8
16
32
Ethoxzolamide (µM)
Control cells (no virus)
Virus-challenged cells
F
SARS-CoV-2 WK-521
Virus load (TCID$_{50}$/ml)
1000000
100000
10000
1000
100
10
N.D.
0.25
0.5
1
2
4
8
16
32
64
Ethoxzolamide (µM)
G
SARS-CoV-2 delta strain
Virus load (TCID$_{50}$/ml)
1000000
100000
10000
1000
100
10
N.D.
0.25
0.5
1
2
4
8
16
32
64
Ethoxzolamide (µM)

Figure 7. LIGHTHOUSE-based drug repurposing for COVID-19

(A) Enrichment analysis of the top 500 potential statin targets identified in Figure 6 for COVID-19–associated gene sets. Minus $\log_{10}$-transformed q values are shown.

(B) 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.

(C) Vero-TMPRSS2 cells were infected with wild-type SARS-CoV-2 at a multiplicity of infection (MOI) of 0.0001, cultured for 64 h in the presence of the indicated concentrations of ethoxzolamide, and subjected to immunocytofluorescence analysis with antibodies to SARS-CoV-2 N protein (green). Nuclei were stained with Hoechst 33342 (blue). Scale bars, 100 μm.

(D and E) Vero-TMPRSS2 cells challenged with wild-type (WK-521) or delta strains of SARS-CoV-2, respectively, were cultured in the presence of various concentrations of ethoxzolamide for 3 days and then subjected to the MTT assay of cell viability. Nonchallenged cells were examined as a control. Data are means ± SD for independent experiments each performed in duplicate.

(F and G) Effect of ethoxzolamide on the SARS-CoV-2 load in culture supernatants of Vero-TMPRSS2 cells challenged with wild-type (WK-521) or delta strains of the virus, respectively. Data are from independent experiments, with the graph line connecting mean values. $TCID_{50}$, median tissue culture infectious dose; N.D., not detected.

See also Figures S9–S12 and Tables S7–S9.

lipid-related pathways such as atherosclerosis and fatty liver, ''pathways in cancer'' was one of the most enriched KEGG pathways (Figure 6C), consistent with previous findings (Ahern et al., 2014; Alfaqih et al., 2017; Matusewicz et al., 2015; Mullen et al., 2016). Potential targets of statins for cancer treatment identified by LIGHTHOUSE included STAT3, CCND1, AKT1, and CCL2 (Figure 6D).

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 (Figure 6E), consistent with the liver being the primary site of statin metabolism and interstitial pneumonia being one of the most severe side effects of statins (Momo et al., 2018). 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 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 (Hu et al., 2021). We noticed that genes whose expression is up-regulated after SARS-CoV-2 infection (Blanco-Melo et al., 2020; Riva et al., 2020; Wyler et al., 2021) were enriched in the list of potential statin targets identified by LIGHTHOUSE (Figure 7A). Indeed, previous studies have shown that statins prevent exacerbation of COVID-19 (Gupta et al., 2021; Zhang et al., 2020). 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 (Walls et al., 2020), and the top drug candidate, ethoxzolamide, was selected for validation analysis (Figure 7B). Immunocytofluorescence analysis revealed that ethoxzolamide blocks proliferation of SARS-CoV-2 in Vero-TMPRSS2 cells (Figure 7C). Furthermore, ethoxzolamide was effective against not only the wild-type virus but also the alpha, beta, gamma, and delta variants. It thus rescued virus-challenged cells in a concentration-dependent manner without affecting non-infected cells (median cytotoxicity concentration of >50 μM) (Figures 7D, 7E, and S9; Table S7), and it reduced the virus load present in the culture supernatant of the cells (Figures 7F, 7G, and S10). Ethoxzolamide is approved for the treatment of seizures and glaucoma (Ghorai et al., 2020; Pospelov et al., 2021), 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.

Experimental evaluation of the potential inhibitory effect of acetazolamide, which was ranked second in the list of potential ACE2-targeting drugs predicted by LIGHTHOUSE and has a known mechanism of action (carbonic anhydrase inhibitor) similar to that of ethoxzolamide, revealed that acetazolamide did not inhibit SARS-CoV-2 infection, despite its LIGHTHOUSE confidence score for ACE2 being very similar to that of ethoxzolamide (0.837 versus 0.881, respectively). We therefore hypothesized that ethoxzolamide might preferentially target a SARS-CoV-2 protein rather than host ACE2. To test this hypothesis, we computed the scores of ethoxzolamide and acetazolamide for each of the SARS-CoV-2 protein sequences (4789 entries in the UniProt database, including predictions) (Figure S11). The virus-derived protein with the largest difference in confidence scores (ethoxzolamide minus acetazolamide) was spike protein S, and ORF7, a protein involved in host-virus interaction, showed the second-largest difference (Table S8). These results suggested that ethoxzolamide also targets the viral S protein, whereas acetazolamide does not. Molecular docking simulation suggested that ethoxzolamide binds to the interface between ACE2 and S protein and thereby blocks virus entry into host cells (Figure S12).

We tested another 10 drugs and found another 2 compounds that also inhibited SARS-CoV-2 infection (Table S9). This high hit rate of 25% (3 hits in 12 compounds) shows the potential power of LIGHTHOUSE for drug discovery.

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 as well as for other AI methodologies.

Existing in silico drug discovery methods can be broadly classified into two categories: SBDD (structure-based drug discovery) and LBDD (ligand-based drug discovery). SBDD requires the 3D structure of the target protein and manual configuration of the binding box. LBDD requires a list of existing compounds with activity data for the target protein. These substantial requirements limit the applicability of AI-powered drug discovery. LIGHTHOUSE is a computational method that eliminates the drawbacks of these two approaches: It does not require the 3D structure of the target protein or the setting of the binding box, and it can be applied to "undruggable" targets for which no inhibitors are currently available. Conventional methods, in particular SBDD, require desktop computers or high-performance cloud computing (such as Amazon EC2) and can calculate only up to 10 CPIs per minute. In contrast, LIGHTHOUSE requires only a simple laptop computer and is able to calculate >1000 CPIs per minute. The time and cost savings of LIGHTHOUSE are therefore several orders of magnitude relative to conventional docking simulation techniques.

We have applied LIGHTHOUSE to attractive targets for various diseases, including cancer, bacterial infection, metabolic diseases, and COVID-19. We have presented three examples that show the effectiveness of LIGHTHOUSE with experimental validation. The actual targets of identified compounds require further biological investigation, especially in the case of pyridoxal 5′-phosphate and PBPs. As for ethoxzolamide, we attempted to obtain clinical evidence in support of its effectiveness against COVID-19. However, ethoxzolamide is an old drug and is essentially no longer prescribed as a result of the development of more potent agents such as acetazolamide. We were therefore not able to find data to address whether COVID-19 patients taking ethoxzolamide have a better clinical outcome. We were also not able to decipher the mechanism of action of ethoxzolamide, but our results suggest that the drug might block the interaction between ACE2 and the viral S protein (Figure S12). The beta, gamma, and delta variants of SARS-CoV-2 appeared to have a higher threshold for drug effectiveness compared with the original and alpha strains (Figures 7F, 7G, and S10). However, when the drug concentration exceeds several micro molar, the virus load sharply drops to the level of undetected (please bear in mind that the y-axis is $\log_{10}$ transformed). The important point with regard to our manuscript is that the virus-challenged cells could be rescued by applying ethoxzolamide, thereby providing experiential validation of the prediction of LIGHTHOUSE.

As for COVID-19, there are many published AI systems to help clinicians. For example, various imaging data (X-ray, CT, MRI, etc.), blood tests, age, gender, region, etc., are integrated to predict the severity of COVID-19 and the occurrence of various symptoms, which definitely help physicians to choose the appropriate approved drugs for that patient (Jamshidi et al., 2020). LIGHTHOUSE differs from these AI models in that it discovers drugs on the basis of the amino acid sequence of a target protein, not on that of clinical information. Furthermore, LIGHTHOUSE is able to identify potential drugs not only for COVID-19 but also for other diseases including cancer, bacterial infection, and lifestyle-related conditions.

The hit compounds themselves identified in the present study are not sufficient for immediate clinical use. An additional potential application of LIGHTHOUSE is drug optimization. One promising method to support such optimization is to apply LIGHTHOUSE and either reinforcement learning (Pereira et al., 2021) (Figure S13) 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) (Lewell et al., 1998). LIGHTHOUSE can score the generated virtual compounds and help to narrow down the candidates with better scores than the original hit compound. Selected candidates can then be synthesized in collaboration with organic chemists and their effects tested. Given the recent success of the MH approach in various life science fields (Biswas et al., 2021; Chen et al., 2021), LIGHTHOUSE should be able to facilitate the optimization of drug candidates by iterating these steps.

LIGHTHOUSE is not limited to compounds for drug repurposing; it can be applied to any compound for which an SMILES representation is available. LIGHTHOUSE facilitates the discovery of compounds that have the potential to interact with targets for which no binding agents have previously been identified. For instance, PPAT has had no known inhibitors and has therefore been categorized as undruggable. In addition, the 3D structure of PPAT has not been reported, with the result that conventional docking simulation methods cannot be applied to discover potential inhibitors. With the use of the amino acid sequence of PPAT alone, LIGHTHOUSE was able to predict potential inhibitors, one of which, compound ZINC8551105, was indeed shown to inhibit PPAT activity. PPAT is therefore no longer an undruggable target. Furthermore, with the use of a hit compound as a starting point, LIGHTHOUSE can be applied to drug optimization. We have calculated the LIGHTHOUSE scores for thousands of analogs of hit compounds that interact with PPAT, and have thereby discovered several PPAT inhibitors that are more potent than the original compounds (manuscript in preparation). LIGHTHOUSE can thus be applied to *de novo* drug discovery and drug optimization as well as to drug repurposing.

Open Targets is a tool that allows data mining on the basis of the function of a gene of interest and a summary of genome-wide association study (GWAS) data available for the relation of the gene to disease (Ochoa et al., 2022). PandaOmics automatically executes differentially expressed gene (DEG) searches and network analysis on the basis of publicly available data from The Cancer Genome Atlas and GEO databases or user-uploaded data. In addition, the AI system of PandaOmics displays and ranks genes that are the most successful therapeutic targets for a disease (Zhavoronkov et al., 2019). Although both of these methods are relatively effective for target identification, they are not designed to identify medications for the target genes. The identification of promising therapeutic target genes by Open Targets and PandaOmics can thus be followed by the prediction of lead compounds for these targets by LIGHTHOUSE.

An important improvement will be to reduce the number of false positives, which may necessitate taking advantage of CPIs that have not been experimentally explored with the use of methods such as collaborative filtering. The next step for more sophisticated AI-dependent drug discovery and its clinical application will be the use of huge-scale and robust protein-chemical binding (and nonbinding) data, with the recent introduction of academic journals that specialize in big data reflecting the growth of this field. Moreover, development of improved approaches to the handling of imbalanced data (Park et al., 2021) is another active research field in computer science. These advances in both data and methodology should allow more reliable modeling of physical interactions and further facilitate drug discovery in the coming years. For *de novo* drug design, it will be necessary to develop an approach to efficiently explore the huge chemical space (10^{60} compounds), and the development of algorithms using Bayesian optimization with appropriate acquisition function as well as chemical generators (Figure S13) is desirable.

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 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.

Limitations of the study

The biggest 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. This confidence score therefore does not necessarily reflect actual interaction for a given protein-chemical pair, and well-studied molecules are thus prone to score higher than others. On the other hand, low confidence scores do not necessarily mean that the protein and chemical in question do not interact. We modeled this score because current biologically determined CPI datasets contain fewer CPI pairs. In addition, as a result of publication bias and biological experimental conditions, it is sometimes difficult to tell whether a protein and chemical do not interact, whether they did not bind under the specific assay condition, or whether they were not tested. This drawback of LIGHTHOUSE 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. Furthermore, we also modeled IC_{50} data from BindingDB, which is derived from actual bioassays. We found that the predicted value and observed value correlate well in the BindingDB test dataset (Figure S6). By combining the confidence score (from STITCH) and the interaction score (from BindingDB) provided by LIGHTHOUSE, we were able to discover a clinically approved drug that was able to block SARS-CoV-2 infection.

Another potential limitation is that the performance for STITCH data might be overestimated because we downsampled the original imbalanced data (Table S1). There are several ways to tackle imbalanced data, including downsampling, oversampling, and more complex sampling methods such as SMOTE (Chawla et al., 2011). Each of these approaches has potential limitations, however (Kulkarni et al., 2021; Yu and Zhou, 2021). In the present study, we adopted downsampling, as in a previous study (Tsubaki et al., 2019), so the performance for STITCH data may be overestimated when compared with use of the entire STITCH data. Furthermore, LIGHTHOUSE can in principle handle only low molecular weight compounds, and it is therefore challenging to apply it directly to the design of nucleic acid, which is gaining popularity as biopharmaceuticals and biosensors.

Despite these limitations, 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 (Bepler and Berger, 2021) and graph-based chemoinformatics approaches, LIGHTHOUSE has the potential to be a cornerstone of drug discovery. It is also of note that we split the STITCH dataset once, given that a previous study (Tsubaki et al., 2019) showed it could obtain high performance with a relatively small range of hyperparameter tuning. In-depth hyper-parameter tuning with cross-validation may further boost the performance of the model.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105314.

ACKNOWLEDGMENTS

This work was supported by KAKENHI grants from the Japan Society for the Promotion of Science (JSPS) to H. Shimizu (JP21K17856), M.K. (JP21K15068), and K.I.N. (JP18H05215), as well as by a grant from the Japan Agency for Medical Research and Development (AMED) to K.I.N. (21wm0425002). H. Shimizu was also supported by postdoctoral fellowships fromTakeda 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. S. 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.S. performed COVID-19 infection analyses. M.M. contributed to discussion. K.I.N. supervised the study and edited the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 15, 2022
Revised: August 8, 2022
Accepted: October 5, 2022
Published: November 18, 2022

REFERENCES

Ahern, T.P., Lash, T.L., Damkier, P., Christiansen, P.M., and Cronin-Fenton, D.P. (2014). Statins and breast cancer prognosis: evidence and opportunities. Lancet Oncol. *15*, e461–e468. https://doi.org/10.1016/S1470-2045(14)70119-6.

Alfaqih, M.A., Allott, E.H., Hamilton, R.J., Freeman, M.R., and Freedland, S.J. (2017). The current evidence on statin use and prostate cancer prevention: are we there yet? Nat. Rev. Urol. *14*, 107–119. https://doi.org/10.1038/nrurol.2016.199.

Bepler, T., and Berger, B. (2021). Learning the protein language: evolution, structure, and function. Cell Syst. *12*, 654–669.e3. https://doi.org/10.1016/j.cels.2021.05.017.

Berkers, C.R., Verdoes, M., Lichtman, E., Fiebiger, E., Kessler, B.M., Anderson, K.C., Ploegh, H.L., Ovaa, H., and Galardy, P.J. (2005). Activity probe for in vivo profiling of the specificity of proteasome inhibitor bortezomib. Nat. Methods *2*, 357–362. https://doi.org/10.1038/sj.leu.2404414.

Biswas, S., Khimulya, G., Alley, E.C., Esvelt, K.M., and Church, G.M. (2021). Low-N protein engineering with data-efficient deep learning. Nat. Methods *18*, 389–396. https://doi.org/10.1038/s41592-021-01100-y.

Blanco-Melo, D., Nilsson-Payant, B.E., Liu, W.C., Uhl, S., Hoagland, D., Møller, R., Jordan, T.X., Oishi, K., Panis, M., Sachs, D., et al. (2020). Imbalanced host response to SARS-CoV-2 drivest development of COVID-19. Cell *181*, 1036–1045.e9. https://doi.org/10.1016/j.cell.2020.04.026.

Borroni, R., Di Blasio, A.M., Gaffuri, B., Santorsola, R., Busacca, M., Viganò, P., and Vignali, M. (2000). Expression of GnRH receptor gene in human ectopic endometrial cells and inhibition of their proliferation by leuprolide acetate. Mol. Cell. Endocrinol. *159*, 37–43. https://doi.org/10.1016/s0303-7207(99)00199-9.

Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2011). SMOTE: synthetic minority over-sampling technique. Preprint at arXiv. https://doi.org/10.48550/arXiv.1106.1813.

Chen, L., Cruz, A., Ramsey, S., Dickson, C.J., Duca, J.S., Hornak, V., Koes, D.R., and Kurtzman, T. (2019). Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. PLoS One *14*, e0220113. https://doi.org/10.1371/journal.pone.0220113.

Chen, X., Zhou, J., Zhang, R., Wong, A.K., Park, C.Y., Theesfeld, C.L., and Troyanskaya, O.G. (2021). Tissue-specific enhancer functional networks for associating distal regulatory regions to disease. Cell Syst. *12*, 353–362.e6. https://doi.org/10.1016/j.cels.2021.02.002.

Deacon, C.F. (2020). Dipeptidyl peptidase 4 inhibitors in the treatment of type 2 diabetes mellitus. Nat. Rev. Endocrinol. *16*, 642–653. https://doi.org/10.1038/s41574-020-0399-8.

Dobson, C.M. (2004). Chemical space and biology. Nature *432*, 824–828. https://doi.org/10.1038/nature03192.

Ghorai, S., Pulya, S., Ghosh, K., Panda, P., Ghosh, B., and Gayen, S. (2020). Structure-activity relationship of human carbonic anhydrase-II inhibitors: detailed insight for future development as anti-glaucoma agents. Bioorg. Chem. *95*, 103557. https://doi.org/10.1016/j.bioorg.2019.103557.

Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., and Dahl, G.E. (2017). Neural message passing for quantum chemistry. Preprint at arXiv. https://doi.org/10.48550/arXiv.1704.01212.

Gilson, M.K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., and Chong, J. (2016). BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. *44*, D1045–D1053. https://doi.org/10.1093/nar/gkv1072.

Greybush, N.J., Pacheco-Peña, V., Engheta, N., Murray, C.B., and Kagan, C.R. (2019). Plasmonic optical and chiroptical response of self-assembled Au nanorod equilateral trimers. ACS Nano *13*, 1617–1624. https://doi.org/10.1021/acsnano.8b07619.

Gupta, A., Madhavan, M.V., Poterucha, T.J., DeFilippis, E.M., Hennessey, J.A., Redfors, B., Eckhardt, C., Bikdeli, B., Platt, J., Nalbandian, A., et al. (2021). Association between antecedent statin use and decreased mortality in hospitalized patients with COVID-19. Nat. Commun. *12*, 1325. https://doi.org/10.1038/s41467-021-21553-1.

Hansen, L.K., and Salamon, P. (1990). Neural network ensembles. IEEE Trans. Pattern Anal. Mach. Intell. *12*, 993–1001. https://doi.org/10.1109/34.58871.

Hsin, K.Y., Matsuoka, Y., Asai, Y., Kamiyoshi, K., Watanabe, T., Kawaoka, Y., and Kitano, H. (2016). systemsDock: a web server for network pharmacology-based prediction and analysis. Nucleic Acids Res. *44*, W507–W513. https://doi.org/10.1093/nar/gkw335.

Hu, B., Guo, H., Zhou, P., and Shi, Z.L. (2021). Characteristics of SARS-CoV-2 and COVID-19. Nat. Rev. Microbiol. *19*, 141–154. https://doi.org/10.1038/s41579-020-00459-7.

Huang, K., Fu, T., Glass, L.M., Zitnik, M., Xiao, C., and Sun, J. (2021). DeepPurpose: a deep learning library for drug-target interaction prediction. Bioinformatics *36*, 5545–5547. https://doi.org/10.1093/bioinformatics/btaa1005.

Jabbour, E., Kantarjian, H., Jones, D., Breeden, M., Garcia-Manero, G., O'Brien, S., Ravandi, F., Borthakur, G., and Cortes, J. (2008). Characteristics and outcomes of patients with chronic myeloid leukemia and T315I mutation following failure of imatinib mesylate therapy. Blood *112*, 53–55. https://doi.org/10.1182/blood-2007-11-123950.

Jamshidi, M.B., Lalbakhsh, A., Talla, J., Peroutka, Z., Hadjilooei, F., Lalbakhsh, P., Jamshidi, M., Spada, L.L., Mirmozafari, M., Dehghani, M., et al. (2020). Artificial intelligence and COVID-19: deep learning approaches for diagnosis and treatment. IEEE Access *8*, 109581–109595. https://doi.org/10.1109/ACCESS.2020.3001973.

Jayatunga, M.K.P., Xie, W., Ruder, L., Schulze, U., and Meier, C. (2022). AI in small-molecule drug discovery: a coming wave? Nat. Rev. Drug Discov. *21*, 175–176. https://doi.org/10.1038/d41573-022-00025-1.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., and Tanabe, M. (2021). KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. *49*, D545–D551. https://doi.org/10.1093/nar/gkaa970.

Kirkpatrick, P. (2022). Artificial Intelligence Makes a Splash in Small-Molecule Drug Discovery. https://www.nature.com/articles/d43747-022-00104-7.

Klein, M.E., Parvez, M.M., and Shin, J.G. (2017). Clinical implementation of pharmacogenomics for personalized precision medicine: barriers and solutions. J. Pharm. Sci. *106*, 2368–2379. https://doi.org/10.1016/j.xphs.2017.04.051.

Knudsen, L.B., and Lau, J. (2019). The discovery and development of liraglutide and semaglutide. Front. Endocrinol. *10*, 155. https://doi.org/10.3389/fendo.2019.00155.

Kodama, M., Oshikawa, K., Shimizu, H., Yoshioka, S., Takahashi, M., Izumi, Y., Bamba, T., Tateishi, C., Tomonaga, T., Matsumoto, M., et al. (2020). A shift in glutamine nitrogen metabolism contributes to the malignant progression of cancer. Nat. Commun. *11*, 1320. https://doi.org/10.1038/s41467-020-15136-9.

Kulkarni, A., Chong, D., and Batarseh, F.A. (2021). Foundations of data imbalance and solutions for a data democracy. Preprint at arXiv. https://doi.org/10.48550/arXiv.2108.00071.

Lewell, X.Q., Judd, D.B., Watson, S.P., and Hann, M.M. (1998). 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. https://doi.org/10.1021/ci970429i.

Li, Q., Han, L., Zhang, B., Zhou, J., and Zhang, H. (2016). Synthesis and biological evaluation of triazole based uracil derivatives as novel DPP-4 inhibitors. Org. Biomol. Chem. *14*, 9598–9611. https://doi.org/10.1039/c6ob01818a.

Liu, H., Sun, J., Guan, J., Zheng, J., and Zhou, S. (2015). Improving compound–protein interaction prediction by building up highly credible negative samples. Bioinformatics *31*, i221–i229. https://doi.org/10.1093/bioinformatics/btv256.

Macheboeuf, P., Contreras-Martel, C., Job, V., Dideberg, O., and Dessen, A. (2006). Penicillin binding proteins: key players in bacterial cell cycle and drug resistance processes. FEMS Microbiol. Rev. *30*, 673–691. https://doi.org/10.1111/j.1574-6976.2006.00024.x.

Matusewicz, L., Meissner, J., Toporkiewicz, M., and Sikorski, A.F. (2015). The effect of statins on cancer cells—review. Tumour Biol. *36*, 4889–4904. https://doi.org/10.1007/s13277-015-3551-7.

Mei, Z., Liang, M., Li, L., Zhang, Y., Wang, Q., and Yang, W. (2017). 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. https://doi.org/10.1002/ijc.30526.

Mohammadi Estakhri, N., Edwards, B., and Engheta, N. (2019). Inverse-designed metastructures that solve equations. Science *363*, 1333–1338. https://doi.org/10.1126/science.aaw2498.

Momo, K., Takagi, A., Miyaji, A., and Koinuma, M. (2018). 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. https://doi.org/10.1016/j.pupt.2018.04.003.

Moussa, N., Hassan, A., and Gharaghani, S. (2021). Pharmacophore model, docking, QSAR, and molecular dynamics simulation studies of substituted cyclic imides and herbal medicines as COX-2 inhibitors. Heliyon *7*, e06605. https://doi.org/10.1016/j.heliyon.2021.e06605.

Mullen, P.J., Yu, R., Longo, J., Archer, M.C., and Penn, L.Z. (2016). The interplay between cell signalling and the mevalonate pathway in cancer. Nat. Rev. Cancer *16*, 718–731. https://doi.org/10.1038/nrc.2016.76.

Muttenthaler, M., King, G.F., Adams, D.J., and Alewood, P.F. (2021). Trends in peptide drug discovery. Nat. Rev. Drug Discov. *20*, 309–325. https://doi.org/10.1038/s41573-020-00135-8.

Ochoa, D., Karim, M., Ghoussaini, M., Hulcoop, D.G., McDonagh, E.M., and Dunham, I. (2022). Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. Nat. Rev. Drug Discov. *21*, 551. https://doi.org/10.1038/d41573-022-00120-3.

Öztürk, H., Özgür, A., and Ozkirimli, E. (2018). DeepDTA: deep drug-target binding affinity prediction. Bioinformatics *34*, i821–i829. https://doi.org/10.1093/bioinformatics/bty593.

Park, S., Lim, J., Jeon, Y., and Choi, J.Y. (2021). Influence-balanced loss for imbalanced visual classification. Preprint at arXiv. https://doi.org/10.48550/arXiv.2110.02444.

Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., and Tekade, R.K. (2021). Artificial intelligence in drug discovery and development. Drug Discov. Today *26*, 80–93. https://doi.org/10.1016/j.drudis.2020.10.010.

Pereira, T., Abbasi, M., Ribeiro, B., and Arrais, J.P. (2021). Diversity oriented deep reinforcement learning for targeted molecule generation. J. Cheminform. *13*, 21. https://doi.org/10.1186/s13321-021-00498-z.

Pospelov, A.S., Ala-Kurikka, T., Kurki, S., Voipio, J., and Kaila, K. (2021). Carbonic anhydrase inhibitors suppress seizures in a rat model of birth asphyxia. Epilepsia *62*, 1971–1984. https://doi.org/10.1111/epi.16963.

Pries, V., Nöcker, C., Khan, D., Johnen, P., Hong, Z., Tripathi, A., Keller, A.L., Fitz, M., Perruccio, F., Filipuzzi, I., et al. (2018). Target identification and mechanism of action of picolinamide and benzamide chemotypes with antifungal properties. Cell Chem. Biol. *25*, 279–290.e7. https://doi.org/10.1016/j.chembiol.2017.12.007.

Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., and Koes, D.R. (2017). Protein-ligand scoring with convolutional neural networks. J. Chem. Inf. Model. *57*, 942–957. https://doi.org/10.1021/acs.jcim.6b00740.

Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine learning in medicine. N. Engl. J. Med. *380*, 1347–1358. https://doi.org/10.1056/NEJMra1814259.

Reczko, M., and Bohr, H. (1994). The DEF data base of sequence based protein fold class predictions. Nucleic Acids Res. 22, 3616–3619.

Riva, L., Yuan, S., Yin, X., Martin-Sancho, L., Matsunaga, N., Pache, L., Burgstaller-Muehlbacher, S., De Jesus, P.D., Teriete, P., Hull, M.V., et al. (2020). Discovery of SARS-CoV-2 antiviral drugs through large-scale compound repurposing. Nature *586*, 113–119. https://doi.org/10.1038/s41586-020-2577-1.

Sasaki, M., Uemura, K., Sato, A., Toba, S., Sanaki, T., Maenaka, K., Hall, W.W., Orba, Y., and Sawa, H. (2021). 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. https://doi.org/10.1371/journal.ppat.1009233.

Seifert, R., Strasser, A., Schneider, E.H., Neumann, D., Dove, S., and Buschauer, A. (2013).

Molecular and cellular analysis of human histamine receptor subtypes. Trends Pharmacol. Sci. *34*, 33–58. https://doi.org/10.1016/j.tips.2012.11.001.

Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., Geng, Q., Auerbach, A., and Li, F. (2020). Structural basis of receptor recognition by SARS-CoV-2. Nature *581*, 221–224. https://doi.org/10.1038/s41586-020-2179-y.

Shimizu, H., and Nakayama, K.I. (2020). Artificial intelligence in oncology. Cancer Sci. *111*, 1452–1460. https://doi.org/10.1111/cas.14377.

Shimizu, H., and Nakayama, K.I. (2021). A universal molecular prognostic score for gastrointestinal tumors. NPJ Genom. Med. *6*, 6. https://doi.org/10.1038/s41525-021-00172-1.

Sterling, T., and Irwin, J.J. (2015). Zinc 15—ligand discovery for everyone. J. Chem. Inf. Model. *55*, 2324–2337. https://doi.org/10.1021/acs.jcim.5b00559.

Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N.M., MacNair, C.R., French, S., Carfrae, L.A., Bloom-Ackermann, Z., et al. (2020). A deep learning approach to antibiotic discovery. Cell *180*, 688–702.e13. https://doi.org/10.1016/j.cell.2020.01.021.

Szklarczyk, D., Santos, A., von Mering, C., Jensen, L.J., Bork, P., and Kuhn, M. (2016). Stitch 5: augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Res. *44*, D380–D384. https://doi.org/10.1093/nar/gkv1277.

Tooke, C.L., Hinchliffe, P., Bragginton, E.C., Colenso, C.K., Hirvonen, V.H.A., Takebayashi, Y., and Spencer, J. (2019). β-Lactamases and β-lactamase inhibitors in the 21st century. J. Mol. Biol. *431*, 3472–3500. https://doi.org/10.1016/j.jmb.2019.04.002.

Trott, O., and Olson, A.J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. J. Comput. Chem. *31*, 455–461. https://doi.org/10.1002/jcc.21334.

Tsubaki, M., Tomii, K., and Sese, J. (2019). Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics *35*, 309–318. https://doi.org/10.1093/bioinformatics/bty535.

UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. *49*, D480–D489. https://doi.org/10.1093/nar/gkaa1100.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. Preprint at arXiv. https://doi.org/10.48550/arXiv.1706.03762.

Wallach, I., Dzamba, M., and Heifets, A. (2015). AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. Preprint at arXiv. https://doi.org/10.48550/arXiv.1510.02855.

Walls, A.C., Park, Y.J., Tortorici, M.A., Wall, A., McGuire, A.T., and Veesler, D. (2020). Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell *181*, 281–292.e6. https://doi.org/10.1016/j.cell.2020.02.058.

Wang, S., Jiang, J.H., Li, R.Y., and Deng, P. (2020). Docking-based virtual screening of TβR1 inhibitors: evaluation of pose prediction and scoring functions. BMC Chem. *14*, 52. https://doi.org/10.1186/s13065-020-00704-3.

Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. *46*, D1074–D1082. https://doi.org/10.1093/nar/gkx1037.

World Health Organization (2021). WHO Model List of Essential Medicines - 22nd List. https://www.who.int/publications/i/item/WHO-MHP-HPS-EML-2021.02.

Wyler, E., Mösbauer, K., Franke, V., Diag, A., Gottula, L.T., Arsiè, R., Klironomos, F., Koppstein, D., Hönzke, K., Ayoub, S., et al. (2021). Transcriptomic profiling of SARS-CoV-2 infected human cell lines identifies HSP90 as target for COVID-19 therapy. iScience *24*, 102151. https://doi.org/10.1016/j.isci.2021.102151.

Yu, L., and Zhou, N. (2021). Survey of imbalanced data methodologies. Preprint at arXiv. https://doi.org/10.48550/arXiv.2104.02240.

Zhang, X.J., Qin, J.J., Cheng, X., Shen, L., Zhao, Y.C., Yuan, Y., Lei, F., Chen, M.M., Yang, H., Bai, L., et al. (2020). In-hospital use of statins is associated with a reduced risk of mortality among individuals with COVID-19. Cell Metab. *32*, 176–187.e4. https://doi.org/10.1016/j.cmet.2020.06.015.

Zhavoronkov, A., Ivanenkov, Y.A., Aliper, A., Veselov, M.S., Aladinskiy, V.A., Aladinskaya, A.V., Terentiev, V.A., Polykovskiy, D.A., Kuznetsov, M.D., Asadulaev, A., et al. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat. Biotechnol. *37*, 1038–1040. https://doi.org/10.1038/s41587-019-0224-x.

Zheng, Y., Ley, S.H., and Hu, F.B. (2018). Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nat. Rev. Endocrinol. *14*, 88–98. https://doi.org/10.1038/nrendo.2017.151.

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat. Commun. *10*, 1523. https://doi.org/10.1038/s41467-019-09234-6.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Python code for LIGHTHOUSE	This paper	https://github.com/Shimizu-Lab/LIGHTHOUSE
Python 3.7.12	Python Software Foundation	https://www.python.org
R 4.1.3	The Comprehensive R Archive Network	https://cran.r-project.org
JMP Pro 15	JMP Statistical Discovery	https://www.jmp.com/en_us/home.html
ADFRsuite 1.0	The Scripps Research Institute	https://ccsb.scripps.edu/adfr/downloads
AutoDock Vina software 1.2.3	Trott and Olson (2010)	https://vina.scripps.edu

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Keiichi I. Nakayama (nakayak1@bioreg.kyushu-u.ac.jp).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All data used for training were downloaded and are publicity available from STITCH (http://stitch.embl.de) (Szklarczyk et al., 2016) and BindingDB (https://www.bindingdb.org/bind/index.jsp) (Gilson et al., 2016) web servers. SMILES representations for small-molecule compounds were downloaded from PubChem (https://pubchem.ncbi.nlm.nih.gov) or ZINC15 (https://zinc15.docking.org) (Sterling and Irwin, 2015). Amino acid sequences were obtained from UniProt (https://www.uniprot.org) (UniProt Consortium, 2021). For Figure S11, we obtained all registered proteins associated with SARS-CoV-2 (https://www.uniprot.org/taxonomy/2697049) and filtered out those containing >7000 amino acids. For drug repurposing analyses, we used the KEGG-DRUG database (https://www.genome.jp/kegg/drug) (Kanehisa et al., 2021). DrugBank (https://go.drugbank.com) was used for identification of target proteins for WHO essential drugs. All referenced COVID-19 signatures are available at Coronascape (https://metascape.org/COVID) (Zhou et al., 2019).
- Code for LIGHTHOUSE with pretrained weights together with a notebook reproducing the results presented in this paper is available at https://github.com/Shimizu-Lab/LIGHTHOUSE.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

SARS-CoV-2 assays

Vero-TMPRSS2 cells (Sasaki et al., 2021) 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 (Sasaki et al., 2021). The MTT assay was performed to evaluate cell viability after virus infection also as previously described (Sasaki et al., 2021). In brief, serial two-fold 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 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–Thermo Fisher Scientific, Cat# A32731). Nuclei were stained with Hoechst 33342 (Invitrogen). Fluorescence images were captured with an IX73 fluorescence microscope (Olympus).

PPAT activity assay

Sf21 cells were cultured in Sf-900 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). 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 (Pries et al., 2018) 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 adjusted to 7.0), after which OD_{600} was measured with a GENESYS 30 visible spectrophotometer (Thermo Fisher 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.

METHOD DETAILS

Generation of a dataset for the training phase of LIGHTHOUSE

The compound SMILES strings of the dataset were extracted from the PubChem compound database on the basis of compound names and PubChem compound IDs (CIDs). The protein sequences of the dataset 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 dataset 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 (Table S1). We then randomly separated these data into training (80%), validation (10%), and test (10%) datasets (Figure S1A). 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%) datasets (Figure S2A). Given that IC_{50} values differ widely, we scaled the values by log transformation (Equation 1) and used the transformed values for BindingDB training.

$$IC_{50\ (scaled)} = -\ log_{10}\left(IC_{50}[M] + 10^{-10}\right) \qquad \text{(Equation 1)}$$

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 (Gilmer et al., 2017). 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 (Equations 2 and 3):

$$m_v^{t+1} = \sum_{w \in N(v)} M_t\left(h_v^t\ ,\ h_w^t, e_{vw}\right)\pi \qquad \text{(Equation 2)}$$

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

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 (Öztürk et al., 2018). 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. AAC is an 8420-length vector in which each position corresponds to a sequence of three amino acids (Reczko and Bohr, 1994). Transformer uses a self-attention–based transformer encoder (Vaswani et al., 2017) that operates on the substructure partition fingerprint of proteins. Algorithmically speaking, Transformer follows $O(n^2)$ in computation time and memory, where n is the input size. 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 (Huang et al., 2021).

As for the decoder, we exploited a previously described architecture (Paul et al., 2021). 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) (Shimizu and Nakayama, 2020), $g(x) = \max(0,x)$, as the activation function in the decoder network.

We defined our loss function with MSE (Equation 4):

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

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 dataset) 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 (Figures S1B–S1G). 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 (Figures S2B–S2G), according to the same guidelines. After the training was completed, we finally evaluated the models with the test datasets, which were kept aside during the training and so had not previously been seen by the models.

Linear combination of confidence and interaction scores

We generated a combined score by linear combination of the confidence and interaction scores (confidence + alpha*interaction), where alpha represents the relative weight of the interaction score. We initially set alpha in order to scale the maximum values of the confidence (max 1) and interaction (max 10) scores. We then conducted a grid search with regard to alpha, which ranges from 0.05 to 0.2, and identified the optimal alpha to distinguish between positive and negative data. Finally, the combination score was defined as confidence score plus 0.075*interaction score.

Generation of virtual chemical libraries and prediction by LIGHTHOUSE

We prepared nearly 1 billion purchasable substances, which were downloaded from the ZINC database (Sterling and Irwin, 2015) as of 30 July 2020, for virtual PPAT inhibitor screening. For drug repurposing, we obtained approved drugs from the KEGG-DRUG database (Kanehisa et al., 2021) 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. As for the peptide drugs shown in Figure 2, we converted them as for small-molecule compounds with the use of SMILES.

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 (Zhou et al., 2019).

Molecular docking simulation

The crystal structure of the receptor binding domain of the spike protein of SARS-CoV-2 in complex with ACE2 (Shang et al., 2020) was downloaded from PDB (accession number: 6VW1) and converted to a pdbqt file with ADFRsuite (version 1.0) according to the recommendations of the developers. A grid box was then set with the following parameters in angstroms: center (X, Y, Z) = (80, 0, 180) and dimensions (X, Y, Z) = (30, 30, 30). Docking simulation was performed with the use of AutoDock Vina software (version 1.2.3) (Trott and Olson, 2010) and with an exhaustiveness of 32.

QUANTIFICATION AND STATISTICAL ANALYSIS

A p value of <0.05 was regarded as statistically significant. Statistical tests are indicated in the results or STAR★methods sections as well as in figure legends. We used R software (version 4.1.3) for statistical analysis, with the exception of the IC_{50} calculation presented in Figure 3, for which we used JMP Pro (version 15.1.0).