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] and Keiichi I. Nakayama[1,11*]

[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]Lead contact

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

Abstract

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 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. We envision that LIGHTHOUSE will bring about a paradigm shift in translational medicine, providing a bridge from bench side to bedside.

KEYWORDS

Artificial intelligence, drug discovery, drug repurposing, COVID-19

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 HTS in most cases has been performed with only ~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,[3] 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.

Recent advances in artificial intelligence (AI) have demonstrated its potential in the pharmaceutical industry.[4] 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,[5] many others have performed only computer-based validation without proof-of-concept biomedical experiments.[6–8] 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[9] 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.[10] 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 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[11] 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),[6] Transformer,[12] and AAC (amino acid composition up to 3-mers).[13] 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 (Table S1), which is one of the largest CPI databases[14] and registers compound-protein pairs together with confidence scores. These scores are based on experimental data, evolutionary evidence such as homologous protein

126 and compound relations, and co-occurrence frequencies in literature abstracts
127 (scores range from 0 to 1, with 1 being the most reliable). To avoid overfitting,
128 we randomly divided the overall data into training (80%), validation (10%), and
129 test (10%) data sets (Supplementary Fig. 1A).
130 We fed the network with protein primary structures and chemicals and
131 trained it to output the scores from the STITCH training data set (Fig. 1B). When
132 we trained the three models (CNN, AAC, and Transformer for protein encoders)
133 separately, the mean squared error (MSE) for the validation data was gradually
134 decreased, and the area under the receiver operating characteristic curve
135 (AUROC) was also improved (Supplementary Fig. 1B–1G). These findings
136 indicated that our AI models learned the approximation of the hidden 1D relation
137 underlying the compound-protein pairs without overfitting the training data. We
138 examined the performance of the models with the test data set at the end of the
139 training and (epoch-wise) validation phases, and we discovered that the AUROC
140 for all three models was >0.80 (Supplementary Table 2). These scores are
141 equivalent to or better than those of cutting-edge 3D docking simulations.[15–17] It
142 is of note that our AI models can be applied to proteins for which 3D structural
143 information is not available. We took the harmonic mean of the three scores to
144 define the confidence score (Fig. 1B).
145 We also trained the models to predict scores based on IC_{50} values. For
146 this purpose, we used data from BindingDB,[18] which collects a variety of
147 experimental findings, and we divided the data into training (80%), validation
148 (10%), and test (10%) data sets (Supplementary Fig. 2A). The same architecture
149 was adopted to train the AI models to predict scaled IC_{50} values (Fig. 1C),
150 yielding an interaction score, and we confirmed that the models adequately
151 learned how to predict IC_{50} from amino acid sequence–chemical pairs
152 (Supplementary Fig. 2B–2G). Finally, we assessed the performance of the
153 models with undisclosed test data, finding that they performed well in predicting
154 IC_{50} (Supplementary Table 3).
155
156 **LIGHTHOUSE architecture outperforms state-of-the-art methods**
157 We next compared the performance of LIGHTHOUSE with that of similar
158 existing methods. To ensure a fair comparison, we used DUD-E data as were
159 used in previous studies.[7,19] In brief, we randomly split DUD-E data (102 target
160 proteins) into training (72 proteins) and test (30 proteins) data and then trained
161 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 (Supplementary Fig. 3A). 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 data set of DUD-E—the training samples comprise 22,886 active (positive) and 22,886 decoy (negative) samples—as in a previous study.[7] LIGHTHOUSE yielded an AUROC for the DUD-E test data of 0.956 (Supplementary Fig. 3B), which was higher than the values produced by state-of-the-art methods including 3D-CNN,[20] AtomNet,[19] and a graph-based deep learning method proposed by Tsubaki et al.[7] (Supplementary Fig. 3C).

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.[7] Both of these data sets were generated previously.[21] 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 (Supplementary Table 4). 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 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 (Figure 2A). Receiver operating characteristic (ROC) curve analysis showed that the two data sets 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" data set. Despite the presence of such false positives, LIGHTHOUSE

198 proved to be effective in predicting the degree of association between
199 protein-chemical pairs solely on the basis of protein primary structure.
200 We next validated the effectiveness of LIGHTHOUSE for well-studied
201 compound-protein pairs. LIGHTHOUSE yielded high confidence scores for
202 adrenergic receptors (α1, α2, β1, β2, and β3) and epinephrine (Fig. 2C).
203 Histamine receptors are classified into four subtypes,[22] with HRH1 and HRH2
204 being targets of antiallergy and antiulcer drugs, respectively. LIGHTHOUSE
205 predicted that the HRH1 antagonist fexofenadine would associate to a greater
206 extent with HRH1 than with HRH2, whereas the HRH2 inhibitor famotidine would
207 associate to a greater extent with HRH2 than with HRH1 (Fig. 2D). These results
208 suggested that LIGHTHOUSE is able to accurately discriminate receptor
209 subtype–level differences solely on the basis of amino acid sequences.
210 LIGHTHOUSE also proved informative both for macrocyclic chemicals
211 such as rapamycin, yielding a high confidence score for this drug and
212 mechanistic target of rapamycin (MTOR) (Fig. 2E), as well as for peptide drugs
213 such as bortezomib (used for treatment of multiple myeloma), leuprorelin
214 (hormone-responsive cancers), and semaglutide (type 2 diabetes) (Fig. 2F),
215 yielding high confidence scores for these drugs and their known targets:
216 proteasome subunit PSMB1,[23] gonadotropin-releasing hormone receptor
217 (GNRHR),[24] and glucagon-like peptide–1 (GLP-1) receptor (GLP1R),[25]
218 respectively. Given the rapidly growing demand for peptide drugs,[26]
219 LIGHTHOUSE will prove useful for the development of novel peptide
220 therapeutics for a variety of promising targets.
221 We also applied LIGHTHOUSE to five drugs that were approved by the
222 U.S. Food and Drug Administration (FDA) in 2020 but which had not yet been
223 registered in the STITCH database. LIGHTHOUSE successfully predicted the
224 association between these new drugs and their target proteins (Fig. 2G),
225 indicating the expandability of LIGHTHOUSE to a much larger exploration space
226 than that encompassed by STITCH.
227 Furthermore, we also evaluated IC_{50} data from BindingDB, which is
228 derived from actual bioassays. We found that the predicted value and observed
229 value correlate well in the BindingDB test data set (Supplementary Fig. 4). This
230 series of findings thus demonstrated the ability of LIGHTHOUSE to discover new
231 drugs for a broad spectrum of diseases.
232

233 **LIGHTHOUSE discovers an inhibitor of PPAT, a key metabolic enzyme for**
234 **cancer treatment**
235 We investigated whether LIGHTHOUSE can identify compounds for novel and
236 potentially important therapeutic targets. As such a target, we chose
237 phosphoribosyl pyrophosphate amidotransferase (PPAT), a rate-limiting enzyme
238 in the *de novo* nucleotide synthesis pathway, given that its expression is most
239 correlated among all metabolic enzymes with poor prognosis in various human
240 cancers and that its depletion markedly inhibits tumor growth[27]. Although no
241 PPAT inhibitor has been developed and the 3D structure of the protein has not
242 been solved, we attempted to discover an inhibitor for PPAT by LIGHTHOUSE
243 solely on the basis of its amino acid sequence. We virtually screened ~10^9
244 commercially available compounds in the ZINC database[28] (Supplementary Fig.
245 5). To reduce the calculation time, we adopted a step-by-step application of
246 LIGHTHOUSE (Fig. 3A). The MPNN_CNN model excluded most of the
247 chemicals unrelated to PPAT, with only 2.41% of the starting compounds having
248 a score of >0.5 in this initial screening (Fig. 3B). The selected compounds were
249 then processed by the MPNN_AAC and MPNN_Transformer models, which
250 reduced the number of candidate chemicals to 0.0356% of the initial compounds.
251 We also calculated interaction scores by LIGHTHOUSE and visualized them in a
252 2D plot (Fig. 3C, left). The best candidates would be expected to have high
253 confidence and interaction scores, appearing in the upper right corner of the plot.
254 Indeed, this criterion was met by several well-known drug-target combinations
255 (Fig. 3C, right).
256 Among the >333,000 final compounds, the top candidate PPAT inhibitor
257 with the highest confidence score was ZINC8551105 (riboflavin
258 5′-monophosphate), with a predicted IC_{50} of 1 to 10 μM (Fig. 3D). We performed
259 a biochemical assay to test this prediction and found that riboflavin
260 5′-monophosphate indeed markedly inhibited PPAT activity with an actual IC_{50} of
261 7.5 μM (Fig. 3E). This compound, discovered by LIGHTHOUSE solely on the
262 basis of the PPAT amino acid sequence, is thus a potential lead compound for
263 the development of new therapeutics targeted to a variety of cancers. It is also of
264 note that we tested only this compound, so other top candidates might also
265 inhibit PPAT activity.
266
267 **LIGHTHOUSE identifies an inhibitor of drug-resistant bacterial growth**

268 Bacterial infections pose a clinical problem worldwide, especially in developing
269 countries, and the emergence of drug-resistant bacterial strains as a result of the
270 overuse of antibiotics has exacerbated this problem. β-Lactamase enzymes
271 produced by antibiotic-resistant bacteria[29] target the β-lactam ring of antibiotics
272 of the penicillin family. We therefore applied LIGHTHOUSE to search for
273 antibiotics not dependent on β-lactam structure.
274 LIGHTHOUSE predicted that pyridoxal 5'-phosphate might associate
275 with penicillin binding proteins such as PBP2 (*mrdA*), PBP3 (*ftsI*), and PBP5
276 (*dacA*), all of which are essential for cell wall synthesis in *Escherichia coli*[30] (Fig.
277 4A). This compound indeed suppressed the growth of wild-type *E. coli* strain
278 JM109 in a concentration-dependent manner (Fig. 4B). Importantly, pyridoxal
279 5'-phosphate also markedly inhibited the growth of an ampicillin-resistant *E. coli*
280 transformant that produces β-lactamase (Fig. 4C). These results thus suggested
281 that, even though it was trained with human proteins, LIGHTHOUSE can also be
282 applied to nonhuman (even bacterial) proteins.
283
284 **LIGHTHOUSE informs optimization of lead compounds**
285 Diabetes mellitus is also a serious public health concern, with the number of
286 affected individuals expected to increase markedly in the coming decades.[31]
287 Dipeptidyl peptidase–4 (DPP-4) cleaves and inactivates the incretin hormones
288 GLP-1 and glucose-dependent insulinotropic polypeptide (GIP), and DPP-4
289 inhibitors are a new class of antidiabetes drug.[32] Given that LIGHTHOUSE also
290 predicts interaction scores, we examined whether it might also contribute to the
291 optimization step of drug development. Indeed, LIGHTHOUSE accurately
292 predicted the rank order of potency for several recently identified DPP-4 inhibitor
293 derivatives[33] (Fig. 5A). Furthermore, LIGHTHOUSE predicted that removal of
294 the phosphate group would reduce the inhibitory potency of riboflavin
295 5′-monophosphate for PPAT (Fig. 3E), and this prediction was confirmed correct
296 by the finding that the IC_{50} value was increased from 7.5 to 49.9 μM (Fig. 5B).
297 These data suggested the possibility that LIGHTHOUSE is capable of predicting
298 activity cliffs.
299 LIGHTHOUSE is also able to estimate the effect of point mutations on
300 CPIs. For example, the T315I mutation of ABL1 in leukemia cells reduces the
301 efficacy of imatinib,[34] and LIGHTHOUSE accurately predicted the effect of this
302 mutation (Fig. 5C). LIGHTHOUSE is able to provide such insight from only
303 wild-type amino acid sequences, given the lack of variant information in the

304 original training data set. Our results suggest that LIGHTHOUSE is able to
305 predict the effects of small changes in protein or chemical structure, and that this
306 will be the case even if such variants do not exist in nature.
307
308 **LIGHTHOUSE identifies potential on- and off-targets of given compounds**
309 Opposite to the mode of drug discovery for a given protein, LIGHTHOUSE might
310 also be able to identify proteins as potential on- or off-targets for a given
311 compound. To verify this notion, we examined statins, which are HMG-CoA
312 reductase inhibitors widely administered for the treatment of hyperlipidemia.
313 Epidemiological studies have shown that statins not only lower cholesterol,
314 however, but also have effects on cancer, although the target molecules for
315 these effects have remained unclear.[35] We therefore applied LIGHTHOUSE to
316 three representative statins (atorvastatin, cerivastatin, and fluvastatin) and
317 computed confidence scores for all human protein-coding genes (Fig. 6A and 6B,
318 Supplementary Table 5). We then sorted the genes on the basis of these
319 confidence scores and performed Kyoto Encyclopedia of Genes and Genomes
320 (KEGG) pathway enrichment analysis for the top 500 potential statin targets. In
321 addition to lipid-related pathways such as atherosclerosis and fatty liver,
322 "pathways in cancer" was one of the most enriched KEGG pathways (Fig. 6C),
323 consistent with previous findings.[36–39] Potential targets of statins for cancer
324 treatment identified by LIGHTHOUSE included STAT3, CCND1, AKT1, and
325 CCL2 (Fig. 6D).
326 Given that side effects of drugs often manifest in organs that express
327 target proteins, we hypothesized that LIGHTHOUSE might be able to identify
328 which organs are at risk of damage from a given drug. We performed another
329 enrichment analysis for the same top 500 potential statin target genes to
330 determine which organs or cell types preferentially express these genes. The top
331 three candidates were the liver, adipocytes, and lung (Fig. 6E), consistent with
332 the liver being the primary site of statin metabolism and interstitial pneumonia
333 being one of the most severe side effects of statins.[40] Prediction of potential
334 target proteins for a given drug by LIGHTHOUSE will thus provide insight into
335 which organs warrant close monitoring by physicians during treatment with the
336 drug, especially in first-in-human clinical trials.
337
338 **LIGHTHOUSE identifies novel potential therapeutics for COVID-19**

339 SARS-CoV-2 emerged at the end of 2019 and has caused a pandemic of
340 infectious pulmonary disease, COVID-19.[41] We noticed that genes whose
341 expression is up-regulated after SARS-CoV-2 infection[42–44] were enriched in the
342 list of potential statin targets identified by LIGHTHOUSE (Fig. 7A). Indeed,
343 previous studies have shown that statins prevent exacerbation of COVID-19.[45,46]
344 With this finding that LIGHTHOUSE is also effective for COVID-19 drug
345 discovery, we applied it to the virtual screening of ~10,000 approved drugs,
346 given that drug repurposing may allow faster delivery of effective agents to
347 patients in need. We calculated scores for angiotensin-converting enzyme 2
348 (ACE2), which is targeted by SARS-CoV-2 for infection of host cells,[47] and the
349 top drug candidate, ethoxzolamide, was selected for validation analysis (Fig. 7B).
350 Immunocytofluorescence analysis revealed that ethoxzolamide blocks
351 proliferation of SARS-CoV-2 in Vero-TMPRSS2 cells (Fig. 7C). Furthermore,
352 ethoxzolamide was effective against not only the wild-type virus but also the
353 alpha, beta, gamma, and delta variants. It thus rescued virus-challenged cells in
354 a concentration-dependent manner without affecting noninfected cells (median
355 cytotoxicity concentration > 50 μM) (Fig. 7D and 7E, Supplementary Fig. 6,
356 Supplementary Table 6), and it reduced the virus load present in the culture
357 supernatant of the cells (Fig. 7F and 7G, Supplementary Fig. 7). Ethoxzolamide
358 is approved for the treatment of seizures and glaucoma,[48,49] and its
359 pharmacodynamics are therefore known. It is therefore immediately available for
360 repurposing for the treatment of patients with COVID-19, with its further
361 optimization having the potential to save many lives. We tested another 11 drugs
362 and found another 2 compounds that also inhibited SARS-CoV-2 infection
363 (Supplementary Table 7). This high hit rate of 25% (3 hits in 12 compounds)
364 shows the potential power of LIGHTHOUSE for drug repositioning.
365

366 **Discussion**

367 Although recent advances in biological and medical research have uncovered
368 various proteins as promising therapeutic targets in a variety of diseases, the
369 clinical application of these research findings has been limited because of the
370 difficulty in identifying therapeutic chemicals for these targets in a cost-effective
371 and high-throughput manner. Acquisition of 3D structural data for target proteins
372 has been labor-intensive, and processing of such data requires a huge amount
373 of computer capacity and time, resulting in a delay in the translation of research
374 findings from the laboratory to the clinic. We have now shown that

375 LIGHTHOUSE facilitates the identification, from a vast chemical space, of
376 candidate compounds for given target proteins solely on the basis of the primary
377 structure of these proteins. Furthermore, the AUROC for LIGHTHOUSE is
378 equivalent to or better than that for state-of-the-art 3D docking simulation
379 methods as well as for other AI methodologies.
380 We have applied LIGHTHOUSE to attractive targets for various
381 diseases, including cancer, bacterial infection, metabolic diseases, and
382 COVID-19, with some of the suggested chemicals being determined
383 experimentally to be effective for inhibition of the corresponding targets. The
384 actual targets of identified compounds require further biological investigation,
385 especially in the case of pyridoxal 5'-phosphate and PBPs. As for ethoxzolamide,
386 we attempted to obtain clinical evidence in support of its effectiveness against
387 COVID-19. However, ethoxzolamide is an old drug and is essentially no longer
388 prescribed as a result of the development of more potent agents such as
389 acetazolamide. We were therefore not able to find data to address whether
390 COVID-19 patients taking ethoxzolamide have a better clinical outcome.
391 The hit compounds themselves identified in the present study are not
392 sufficient for immediate clinical use. LIGHTHOUSE can be applied not only for
393 the identification of lead compounds but also for their subsequent optimization,
394 which requires extensive work to identify more potent and specific or less toxic
395 derivatives. One promising method to support such optimization is to apply
396 LIGHTHOUSE and either reinforcement learning[50] (Supplementary Fig. 8) or
397 Metropolis-Hasting (MH) approaches together. Virtual libraries can be generated
398 from identified lead compounds in an intensive manner with the use of
399 sophisticated chemoinformatics algorithms such as RECAP (Retrosynthetic
400 Combinatorial Analysis Procedure).[51] LIGHTHOUSE can score the generated
401 virtual compounds and help to narrow down the candidates with better scores
402 than the original hit compound. Select candidates can then be synthesized in
403 collaboration with organic chemists and their effects tested. Given the recent
404 success of the MH approach in various life science fields,[52,53] LIGHTHOUSE
405 should be able to facilitate the optimization of drug candidates by iterating these
406 steps.
407 The biggest limitation of LIGHTHOUSE is the generation of false
408 positives, which is due in part to the fact that the confidence score provided by
409 STITCH is not based solely on experimental data but also on other factors such
410 as co-occurrence in the literature. This confidence score therefore does not

411 necessarily reflect actual interaction for a given protein-chemical pair, and
412 well-studied molecules are thus prone to score higher than others. On the other
413 hand, low confidence scores do not necessarily mean that the protein and
414 chemical in question do not interact. We modeled this score because current
415 biologically determined CPI data sets contain fewer CPI pairs. In addition, as a
416 result of publication bias and biological experimental conditions, it is sometimes
417 difficult to tell whether a protein and chemical do not interact, whether they did
418 not bind under the specific assay condition, or whether they were not tested.
419 This drawback of LIGHTHOUSE can be mitigated partially by combining the
420 three different models (CNN, AAC, and Transformer). It may also be important to
421 perform a counter–virtual screening to determine whether an identified small
422 molecule reacts specifically with the target protein or whether it scores highly
423 with many proteins. Such an approach has the potential to reduce the number of
424 false positives and provide more accurate guidance. Furthermore, we also
425 modeled IC_{50} data from BindingDB, which is derived from actual bioassays. We
426 found that the predicted value and observed value correlate well in the
427 BindingDB test data set (Supplementary Fig. 4). By combining the confidence
428 score (from STITCH) and the interaction score (from BindingDB) provided by
429 LIGHTHOUSE, we were able to discover a clinically approved drug that was
430 able to block SARS-CoV-2 infection.
431 Another potential limitation is that the performance for STITCH data
432 might be overestimated because we downsampled the original imbalanced data
433 (Supplementary Table 1). There are several ways to tackle imbalanced data,
434 including downsampling, oversampling, and more complex sampling methods
435 such as SMOTE.[54] Each of these approaches has potential limitations,
436 however.[55,56] In the present study, we adopted downsampling, as in a previous
437 study,[7] so the performance for STITCH data may be overestimated when
438 compared with use of the entire STITCH data.
439 Despite these limitations, LIGHTHOUSE proved to be effective for the
440 identification of lead compounds for all conditions tested. It can theoretically be
441 applied to any protein of any organism, and even to proteins that do not exist
442 naturally. This is an advantage over 3D docking simulation methods, which
443 require prior 3D structural knowledge of the protein of interest. LIGHTHOUSE
444 computes and embeds structural information in numerical vectors, which are
445 then readily retrieved by the subsequent decoding module. Given the
446 accelerating development of protein embedding technologies[57] and graph-based

447 chemoinformatics approaches, LIGHTHOUSE has the potential to be a
448 cornerstone of drug discovery. It is also of note that we split the STITCH data set
449 once, given that a previous study[7] showed it could obtain high performance with
450 a relatively small range of hyperparameter tuning. In-depth hyperparameter
451 tuning with cross-validation may further boost the performance of the model.
452 The next step for more sophisticated AI-dependent drug discovery and
453 its clinical application will be the use of huge-scale and robust protein-chemical
454 binding (and nonbinding) data, with the recent introduction of academic journals
455 that specialize in big data reflecting the growth of this field. Moreover,
456 development of improved approaches to the handling of imbalanced data[58] is
457 another active research field in computer science. These advances in both data
458 and methodology should allow more reliable modeling of physical interactions
459 and further facilitate drug discovery in the coming years.
460 In summary, we have developed LIGHTHOUSE as a means to discover
461 promising lead compounds for any target protein irrespective of its 3D structural
462 information. Furthermore, we have demonstrated the power of LIGHTHOUSE by
463 identifying and validating novel therapeutics for various global health concerns
464 including COVID-19. LIGHTHOUSE will serve as a guide for researchers in all
465 areas of biomedicine, paving the way for a wide range of future applications.
466

467 **EXPERIMENTAL PROCEDURES**

468 **Resource availability**

469 ***Lead contact***

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

473 ***Materials availability***

474 This study did not generate new unique reagents.

475 ***Data and code availability***

476 All data used for training are downaloaded and publicity available from STITCH
477 (http://stitch.embl.de)[14] and BindingDB
478 (https://www.bindingdb.org/bind/index.jsp)[18] web servers. SMILES
479 representations for small-molecule compounds were downloaded from
480 PubChem (https://pubchem.ncbi.nlm.nih.gov) or ZINC15
481 (https://zinc15.docking.org).[28] Amino acid sequences were obtained from
482 UniProt (https://www.uniprot.org).[59] For drug repurposing analyses, we used the

483 KEGG-DRUG database (https://www.genome.jp/kegg/drug).[60] All referenced
484 COVID-19 signatures are available at Coronascape
485 (https://metascape.org/COVID).[61]
486 Code for LIGHTHOUSE with pretrained weights together with a
487 notebook reproducing the results presented in this paper is available at
488 https://github.com/Shimizu-Lab/LIGHTHOUSE.
489
490 **Generation of a data set for the training phase of LIGHTHOUSE**
491 The compound SMILES strings of the data set were extracted from the
492 PubChem compound database on the basis of compound names and PubChem
493 compound IDs (CIDs). The protein sequences of the data set were extracted
494 from the UniProt protein database on the basis of gene names/RefSeq
495 accession numbers or the UniProt IDs. We downloaded the protein-chemical link
496 data set of *Homo sapiens* (Taxonomy ID 9606) from the STITCH database
497 (version 5.0). Given that the STITCH score is heavily biased toward 0, we
498 separated the data into nine bins on the basis of the score and stratify-extracted
499 the same number of CPIs (140,000 each), yielding 1,260,000 CPIs
500 (Supplementary Table 1). We then randomly separated these data into training
501 (80%), validation (10%), and test (10%) data sets (Supplementary Fig. 1A). With
502 regard to IC_{50}, we downloaded data from BindingDB, obtained SMILES
503 expressions and amino acid sequences similarly, and again separated the data
504 into training (80%), validation (10%), and test (10%) data sets (Supplementary
505 Fig. 2A). Given that IC_{50} values differ widely, we scaled the values by log
506 transformation (Eq. 1) and used the transformed values for BindingDB training.
507
508 $$IC_{50\,(scaled)} = -log_{10}(IC_{50}[M] + 10^{-10}) \quad (1)$$
509
510 **LIGHTHOUSE architecture and training**
511 The proposed overall model comprises two encoder networks (for chemicals and
512 proteins) and one decoder network. MPNN is a message passing graph neural
513 network that operates on compound molecular graphs.[11] In brief, MPNN
514 conveys latent information among the atoms and edges. The message passing
515 phase runs for t time steps and is defined in terms of message functions M_t and
516 vertex update functions U_t. During this phase, hidden states h_v^t (128 dimensions
517 in our model) at each node in the chemical graph are updated with the incoming
518 messages m_v^{t+1} according to the following equations (Eqs. 2 and 3):

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

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

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.[6] 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.[13] Transformer uses a self-attention–based transformer encoder[12] 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.[12] 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.[8]

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),[62] $g(x) = \max(0,x)$, as the activation function in the decoder network.

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

553 $$MSE = \frac{1}{n}\sum_{i=1}^{n}(P_i - y_i)^2 \quad (4)$$

554

555 where P_i is the LIGHTHOUSE-predicted score for the ith compound-protein pair
556 and Y_i is the true label in the corresponding training data, with a batch size of
557 128. We trained three architectures (MPNN_CNN, MPNN_AAC,
558 MPNN_Transformer) separately for the STITCH and BindingDB training data
559 with the Adam optimizer and a learning rate of 0.001. For evaluation metrics, we
560 used MSE, concordance index, and Pearson correlation as well as AUROC. For
561 every 10 epochs, we compared the current loss (in the validation data set) with
562 that of 10 epochs ago; if the loss was not decreasing, we terminated the training
563 for that model. As a result of this early termination, we trained MPNN_CNN for
564 40 epochs, MPNN_AAC for 70 epochs, and MPNN_Transformer for 100 epochs
565 with regard to the confidence score (Supplementary Fig. 1B–1G). As for the
566 models for the interaction score, we trained MPNN_CNN for 70 epochs,
567 MPNN_AAC for 100 epochs, and MPNN_Transformer for 70 epochs
568 (Supplementary Fig. 2B–2G), according to the same guidelines. After the
569 training was completed, we finally evaluated the models with the test data sets,
570 which were kept aside during the training and so had not previously been seen
571 by the models.
572

573 **Generation of virtual chemical libraries and prediction by LIGHTHOUSE**
574 We prepared nearly 1 billion purchasable substances, which were downloaded
575 from the ZINC database[28] as of 30 July 2020, for virtual PPAT inhibitor screening.
576 For drug repurposing, we obtained approved drugs from the KEGG-DRUG
577 database[60] as of 24 January 2021. For calculation of confidence and interaction
578 scores, we fixed the proteins of interest (PPAT or ACE2) and changed the
579 compounds iteratively, which yielded lists of predicted scores for all the
580 compounds tested. As for the peptide drugs shown in Fig. 2, we converted them
581 as for small-molecule compounds with the use of SMILES.
582

583 **PPAT activity assay**
584 Sf21 cells were cultured in Sf-900 II SFM (Gibco, Cat# 10902-088)
585 supplemented with 10 μM ferric ammonium citrate. They were transfected with a

586 bacmid encoding human PPAT for 64 h, harvested, washed three times with
587 phosphate-buffered saline, and lysed in a solution containing 150 mM NaCl, 25
588 mM Tris-HCl (pH 7.4), 0.5% Triton X-100, and 5 mM EDTA. The lysate was
589 centrifuged at 10,000 × *g* for 6 min at 4°C, and the resulting supernatant (100
590 ng/ml) was incubated for 4 h at 37°C together with 5 mM glutamine (Gibco, Cat#
591 25030-081), 1 mM phosphoribosyl pyrophosphate (Sigma, Cat# P8296), 10 mM
592 $MgCl_2$, 50 mM Tris-HCl (pH 7.4), and various concentrations of riboflavin
593 5′-monophosphate (Sigma, Cat# F2253-10 mg). Enzyme activity was assessed
594 on the basis of glutamate production as measured with a glutamate assay kit
595 (Abcam, Cat# 138883). The IC_{50} value was estimated from biological
596 quadruplicates with a four-parameter logistic model[63] and with the use of JMP
597 pro 15 software (version 15.1.0).
598
599 **Assay of *E. coli* growth**
600 Portions (20 *μ*l) of *E. coli* strain JM109 (1×10^{10} colony-forming units (CFU)/ml)
601 were cultured for various times in 2 ml of 2xYT liquid medium (BD Difco, Cat#
602 244020) containing various concentrations of pyridoxal 5′-phosphate (pH
603 adjusted to 7.0), after which OD_{600} was measured with a GENESYS 30 visible
604 spectrophotometer (ThermoFisher Scientific, Cat# 840-277000). In addition, the
605 JM109 strain was transformed with 1 μg of the pBlueScript II SK+ plasmid
606 (Invitrogen), which harbors an ampicillin resistance gene as a selection marker,
607 and was then spread on LB agar plates containing ampicillin (100 μg/ml) (Wako,
608 Cat# 012-23303) with or without pyridoxal 5′-phosphate (3 mg/ml) and incubated
609 overnight.
610
611 **Virtual identification of statin targets and enrichment analyses**
612 Three representative statins were fixed as chemical inputs, and all human
613 protein-coding genes in the UniProt database were iteratively changed. The
614 harmonic mean of the three confidence scores was calculated as an affinity
615 score for statins, and the human protein-coding genes were sorted on the basis
616 of this score. The resulting top 500 potential targets were then subjected to
617 enrichment analyses with the use of the Metascape web server.[61]
618
619 **SARS-CoV-2 assays**
620 Vero-TMPRSS2 cells[64] were maintained in Dulbecco's modified Eagle's medium
621 (DMEM) supplemented with 10% fetal bovine serum. The WK-521 strain of

622 SARS-CoV-2 (EPI_ISL_408667) as well as the alpha (QK002,
623 EPI_ISL_768526), beta (TY7-501, EPI_ISL_833366), gamma (TY8-612,
624 EPI_ISL_1123289), and delta (TY11-927, EPI_ISL_2158617) variants were
625 obtained from National Institute of Infectious Diseases in Japan. Stocks of these
626 viruses were prepared by inoculation of Vero-TMPRSS2 cell cultures as
627 described previously.[64] The MTT assay was performed to evaluate cell viability
628 after virus infection also as previously described.[64] In brief, serial twofold
629 dilutions of ethoxzolamide in minimum essential medium (MEM) supplemented
630 with 2% fetal bovine serum were added in duplicate to 96-well microplates.
631 Vero-TMPRSS2 cells infected with wild-type or variant SARS-CoV-2 at 4 to 10
632 $TCID_{50}$ (median tissue culture infectious dose) were also added to the plates,
633 which were then incubated at 37°C for 3 days. The viability of the cells was then
634 determined with the MTT assay, and the culture supernatants were harvested
635 for determination of the $TCID_{50}$ value as a measure of viral load. For indirect
636 immunofluorescence analysis, cells infected with wild-type SARS-CoV-2 at a
637 multiplicity of infection (MOI) of 0.0001 were cultured in the presence of various
638 concentrations of ethoxzolamide for 64 h, fixed with 3.7% buffered formaldehyde,
639 permeabilized with 0.05% Triton X-100, and incubated with antibodies to
640 SARS-CoV-2 N protein (GeneTex, Cat# GTX635679). Immune complexes were
641 detected with Alexa Fluor Plus 488–conjugated goat antibodies to rabbit
642 immunoglobulin G (Invitrogen–ThermoFisher Scientific, Cat# A32731). Nuclei
643 were stained with Hoechst 33342 (Invitrogen). Fluorescence images were
644 captured with an IX73 fluorescence microscope (Olympus).
645
646 **SUPPLEMENTAL INFORMATION**
647 Supplemental information can be found onine at....
648
649 **ACKNOWLEDGMENTS**
650 This work was supported by KAKENHI grants from the Japan Society for the
651 Promotion of Science (JSPS) to H. Shimizu (JP21K17856), M.K. (JP21K15068),
652 and K.I.N. (JP18H05215), as well as by a grant from the Japan Agency for
653 Medical Research and Development (AMED) to K.I.N. (21wm0425002). H.
654 Shimizu was also supported by postdoctoral fellowships from Takeda Science
655 Foundation and Mochida Foundation. We thank T. Sawada and R. Wada for
656 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.

DECLARATION OF INTERESTS

The authors declare no competing interests.

670 **REFERENCES**

671 1. Klein M.E., Parvez M.M. Shin, J.-G. Clinical implementation of
672 pharmacogenomics for personalized precision medicine: barriers and solutions.
673 *J. Pharm. Sci.* 2017; **106**: 2368–2379.
674 2. Dobson C. M. Chemical space and biology. *Nature* 2004; **432**: 824–828.
675 3. Jumper J., Evans R., Pritzel A., Green T., Figurnov M., Ronneberger O.,
676 Tunyasuvunakool K., Bates R., Žídek A., Potapenko A. et al. Highly accurate
677 protein structure prediction with AlphaFold. *Nature* 2021; **596**: 583-589.
678 4. Paul D., Sanap G., Shenoy S., Kalyane D., Kalia K., Tekade R.K. Artificial
679 intelligence in drug discovery and development. *Drug Discov. Today* 2021; **26**:
680 80–93.
681 5. Stokes J.M., Yang K., Swanson K., Jin W., Cubillos-Ruiz A., Donghia N.M.,
682 MacNair C.R., French S., Carfrae L.A., Bloom-Ackermann Z. A Deep Learning
683 Approach to Antibiotic Discovery. *Cell* 2020; **180**: 688-702.
684 6. Öztürk H., Özgür A., Ozkirimli E. DeepDTA: deep drug-target binding affinity
685 prediction. *Bioinformatics* 2018; **34**: i821–i829.
686 7. Tsubaki M., Tomii K., Sese J. Compound-protein interaction prediction with
687 end-to-end learning of neural networks for graphs and sequences.
688 *Bioinformatics* 2019; **35**: 309–318.
689 8. Huang K., Fu T., Glass L.M., Zitnik M., Xiao C., Sun J. DeepPurpose: a deep
690 learning library for drug-target interaction prediction. *Bioinformatics* 2021; **36**:
691 5545–5547.
692 9. Chen L., Cruz A., Ramsey S., Dickson C.J., Duca J.S., Hornak V., Koes D.R.,
693 Kurtzman T. Hidden bias in the DUD-E dataset leads to misleading performance
694 of deep learning in structure-based virtual screening. *PLoS One* 2019; **14**:
695 e0220113.
696 10. Hansen L. K., Salamon P. Neural network ensembles. *IEEE Trans. Pattern*
697 *Anal. Mach. Intell.* 1990; **12**: 993–1001.
698 11. Gilmer J., Schoenholz S.S., Riley P.F., Vinyals O., Dahl G.E. Neural
699 message passing for quantum chemistry. *arXiv* (2017), 1704.01212 [cs.LG].
700 12. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N.,
701 Kaiser L., Polosukhin I. Attention is all you need. arXiv (2017), 1706.03762
702 [cs.CL].
703 13. Reczko M., Bohr H. The DEF data base of sequence based protein fold class
704 predictions. *Nucleic Acids Res.* 1994; **17**: 3616–3619.

705 14. Szklarczyk D., Santos A., Von Mering C., Jensen L.J., Bork P., Kuhn M.
706 STITCH 5: augmenting protein-chemical interaction networks with tissue and
707 affinity data. *Nucleic Acids Res.* 2016; **44**: D380–D384.
708 15. Hsin K.-Y., Matsuoka Y., Asai Y., Kamiyoshi K., Watanabe T., Kawaoka Y.,
709 Kitano H. systemsDock: a web server for network pharmacology-based
710 prediction and analysis. *Nucleic Acids Res.* 2016; **44**: W507–W513.
711 16. Wang S., Jiang J.-H., Li R.-Y., Deng, P. Docking-based virtual screening of
712 TβR1 inhibitors: evaluation of pose prediction and scoring functions. *BMC Chem.*
713 2020; **14**: 52.
714 17. Moussa N., Hassan A., Gharaghani S. Pharmacophore model, docking,
715 QSAR, and molecular dynamics simulation studies of substituted cyclic imides
716 and herbal medicines as COX-2 inhibitors. *Heliyon* 2021; **7**: e06605.
717 18. Gilson M.K., Liu T., Baitaluk M., Nicola G., Hwang L., Chong J. BindingDB in
718 2015: a public database for medicinal chemistry, computational chemistry and
719 systems pharmacology. *Nucleic Acids Res.* 2016; **44**: D1045–D1063.
720 19. Wallach I., Dzamba M., Heifets A. AtomNet: A Deep Convolutional Neural
721 Network for Bioactivity Prediction in Structure-based Drug Discovery. *arXiv* 2015,
722 1510.02855 [cs.LG].
723 20. Ragoza M., Hochuli J., Idrobo E., Sunseri J. and Koes D. R. Protein-Ligand
724 Scoring with Convolutional Neural Networks. *J Chem Inf Model.* 2017; **57**:
725 942-957.
726 21. Liu H., Sun J., Guan J., Zheng J., Zhou S. Improving compound–protein
727 interaction prediction by building up highly credible negative samples.
728 *Bioinformatics* 2015; **31**: i221–i229.
729 22. Seifert R., Strasser A., Schneider E.H., Neumann. D., Dove S., Buschauer A.
730 Molecular and cellular analysis of human histamine receptor subtypes. *Trends*
731 *Pharmacol. Sci.* 2013; **34**: 33–58.
732 23. Berkers C.R., Verdoes M., Lichtman E., Fiebiger E., Kessler B.M., Anderson
733 K.C., Ploegh H.L., Ovaa H., Galardy P.J. Activity probe for in vivo profiling of the
734 specificity of proteasome inhibitor bortezomib. *Nat. Methods* 2005; **2**: 357–362.
735 24. Borroni R., Blasio A.M.D., Gaffuri B., Santorsola R., Busacca M., Viganò P.,
736 Vignali M. Expression of GnRH receptor gene in human ectopic endometrial
737 cells and inhibition of their proliferation by leuprolide acetate. *Mol. Cell.*
738 *Endocrinol.* 2000; **159**: 37–43.
739 25. Knudsen L.B, Lau J. The discovery and development of liraglutide and
740 semaglutide. *Front. Endocrinol. (Lausanne)* 2019; **10**: 155.

741 26. Muttenthaler M., King G.F., Adams D.J., Alewood P.F. Trends in peptide
742 drug discovery. *Nat. Rev. Drug Discov.* 2021; **20**: 309–325.
743 27. Kodama M, Oshikawa K, Shimizu H, Yoshioka S, Takahashi M., Izumi Y.,
744 Bamba T., Tateishi C., Tomonaga T., Matsumoto M. et al. A shift in glutamine
745 nitrogen metabolism contributes to the malignant progression of cancer. *Nat.*
746 *Commun.* 2020; **11**:1320.
747 28. Sterling T., Irwin J.J. ZINC 15—ligand discovery for everyone. *J. Chem. Inf.*
748 *Model.* 2015; **55**: 2324–2337.
749 29. Tooke C.L., Hinchliffe P., Bragginton E.C., Colenso C.K., Hirvonen V.H.A.,
750 Takebayashi Y., Spencer J. β-Lactamases and β-lactamase inhibitors in the
751 21st century. *J. Mol. Biol.* 2019; **431**: 3472–3500.
752 30. Macheboeuf P., Contreras-Martel C., Job C., Dideberg O., Dessen A.
753 Penicillin binding proteins: key players in bacterial cell cycle and drug resistance
754 processes. *FEMS Microbiol. Rev.* 2006; **30**: 673–691.
755 31. Zheng Y., Ley S.H., Hu F.B. Global aetiology and epidemiology of type 2
756 diabetes mellitus and its complications. *Nat. Rev. Endocrinol.* 2018; **14**: 88–98.
757 32. Deacon C. F. Dipeptidyl peptidase 4 inhibitors in the treatment of type 2
758 diabetes mellitus. *Nat. Rev. Endocrinol.* 2020; **16**: 642–653.
759 33. Li Q., Han L., Zhang B., Zhou J., Zhang H. Synthesis and biological
760 evaluation of triazole based uracil derivatives as novel DPP-4 inhibitors. *Org.*
761 *Biomol. Chem.* 2016; **14**: 9598–9611.
762 34. Jabbour E., Kantarjian H., Jones D., Breeden M., Garcia-Manero G., O'Brien
763 S., Ravandi F., Borthakur G., Cortes J. Characteristics and outcomes of patients
764 with chronic myeloid leukemia and T315I mutation following failure of imatinib
765 mesylate therapy. *Blood* 2008; **112**: 53–55.
766 35. Mei Z., Liang M., Li L., Zhang Y., Wang Q., Yang W. Effects of statins on
767 cancer mortality and progression: a systematic review and meta-analysis of 95
768 cohorts including 1,111,407 individuals. *Int. J. Cancer* 2017; **140**: 1068–1081.
769 36. Matusewicz L., Meissner J., Toporkiewicz M., Sikorski A.F. The effect of
770 statins on cancer cells—review. *Tumour Biol.* 2015; **36**: 4889–4904.
771 37. Ahern T.P., Lash T.L., Damkier P., Christiansen P.M., Cronin-Fenton D.P.
772 Statins and breast cancer prognosis: evidence and opportunities. *Lancet Oncol.*
773 2014; **15**: e461–e468.
774 38. Mullen P.J., Yu R., Longo J., Archer M.C., Penn L.Z. The interplay between
775 cell signalling and the mevalonate pathway in cancer. *Nat. Rev. Cancer* 2016;
776 **16**: 718–731.

39. 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.* 2017; **14**: 107–119.
40. 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.* 2018; **50**: 88–92.
41. Hu B., Guo H., Zhou P., Shi Z.-L. Characteristics of SARS-CoV-2 and COVID-19. *Nat. Rev. Microbiol.* 2021; **19**: 141–154.
42. 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*. Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell* 2020; **181**: 1036–1045.
43. 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*. Discovery of SARS-CoV-2 antiviral drugs through large-scale compound repurposing. *Nature* 2020; **586**: 113–119.
44. Wyler E., Mösbauer K., Franke V., Diag A., Gottula L.T., Arsiè R., Klironomos F., Koppstein D., Hönzke Katja., Ayoub S. Transcriptomic profiling of SARS-CoV-2 infected human cell lines identifies HSP90 as target for COVID-19 therapy. *iScience* 2021; **24**: 102151.
45. 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*. In-hospital use of statins is associated with a reduced risk of mortality among individuals with COVID-19. *Cell Metab.* 2020; **32**: 176–187.
46. 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*. Association between antecedent statin use and decreased mortality in hospitalized patients with COVID-19. *Nat. Commun.* 2021; **12**: 1325.
47. Walls A.C., Park Y.-J., Tortorici M.A., Wall A., McGuire A.T., Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* 2020; **181**: 281–292.
48. 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* 2021; **62**: 1971-1984.
49. Ghorai S., Pulya S., Ghosh K., Panda P., Ghosh B., Gayen S. Structure-activity relationship of human carbonic anhydrase-II inhibitors: detailed

812 insight for future development as anti-glaucoma agents. *Bioorg. Chem.* 2020;
813 **95**: 103557.
814 50. Pereira T., Abbasi M., Ribeiro B., Arrais J.P. Diversity oriented deep
815 reinforcement learning for targeted molecule generation. *J. Cheminform.* 2021;
816 **13**: 21.
817 51. Lewell X.Q., Judd D.B., Watson S.P., Hann M.M. RECAP—retrosynthetic
818 combinatorial analysis procedure: a powerful new technique for identifying
819 privileged molecular fragments with useful applications in combinatorial
820 chemistry. *J. Chem. Inf. Comput. Sci.* 1998; **38**: 511–522.
821 52. Biswas S., Khimulya G., Alley E.C., Esvelt K.M., Church G.M. Low-N protein
822 engineering with data-efficient deep learning. *Nat. Methods* 2021; **18**: 389–396.
823 53. Chen X., Zhou J., Zhang R., Wong A.K., Park C.Y., Theesfeld C.L.,
824 Troyanskaya O.G. Tissue-specific enhancer functional networks for associating
825 distal regulatory regions to disease. *Cell Syst.* 2021; **12**: 353–362.
826 54. Chawla N.V., Bowyer K.W., Hall L. O., Kegelmeyer W.P. SMOTE: Synthetic
827 Minority Over-sampling Technique. *arXiv:1106.1813 [cs.AI]*
828 55. Yu L, Zhou N. Survey of Imbalanced Data Methodologies. *arXiv:2104.02240*
829 *[stat.ML]*
830 56. Kulkarni A, Chong D, Batarseh F.A. Foundations of data imbalance and
831 solutions for a data democracy. *arXiv:2108.00071 [cs.LG]*
832 57. Bepler T., Berger B. Learning the protein language: evolution, structure, and
833 function. *Cell Syst*. 2021; **12**: 654–669.
834 58. Park S, Lim J, Jeon Y, Choi J.Y. Influence-Balanced Loss for Imbalanced
835 Visual Classification. *arXiv:2110.02444 [cs.CV]*
836 59. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021.
837 *Nucleic Acids Res.* 2021; **49**: D480–D489.
838 60. Kanehisa M., Furumichi M., Sato Y., Ishiguro-Watanabe M., Tanabe M.
839 KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* 2021; **49**:
840 D545–D551.
841 61. Zhou Y., Zhou B., Pache L., Chang M., Khodabakhshi A. H., Tanaseichuk O.,
842 Benner C., Chanda S.K. Metascape provides a biologist-oriented resource for
843 the analysis of systems-level datasets. *Nat. Commun.* 2019; **10**: 1523.
844 62. Shimizu H., Nakayama K.I. Artificial intelligence in oncology. *Cancer Sci.*
845 2020; **111**: 1452–1460.
846 63. Pries V., Nöcker C., Khan D., Johnen P., Hong Z., Tripathi A., Keller A.-L.,
847 Fitz M., Perruccio F., Filipuzzi I. Target identification and mechanism of action of

848 picolinamide and benzamide chemotypes with antifungal properties. *Cell Chem.*
849 *Biol.* 2018; **25**: 279–290.
850 64. Sasaki M., Uemura K., Sato A., Toba S., Sanaki T., Maenaka K., Hall W.W.,
851 Orba Y., Sawa H. SARS-CoV-2 variants with mutations at the S1/S2 cleavage
852 site are generated in vitro during propagation in TMPRSS2-deficient cells. *PLoS*
853 *Pathog*. 2021; **17**: e1009233.
854

855 **FIGURE LEGENDS**

856 **Figure 1. Development of LIGHTHOUSE for discovery of drug candidates**
857 **without 3D structural data**

858 (A) The basic network structure of LIGHTHOUSE consists of encoder and
859 decoder networks. The basic network encodes the amino acid sequence of the
860 protein of interest as numerical vectors by one of three independent methods:
861 CNN, AAC, and Transformer. It also takes the SMILES representation of each
862 small-molecule compound and computes the neural representation with the
863 MPNN algorithm. The network then concatenates the protein and compound
864 representations and calculates a "Score."
865 (B and C) LIGHTHOUSE consists of two modules. Module 1 estimates the
866 association between a given compound-protein pair, and module 2 predicts a
867 scaled IC_{50} value for the pair. In each module, the three different streams of the
868 basic network (MPNN_CNN, MPNN_AAC, and MPNN_Transformer) are used,
869 and the harmonic mean of the three scores is presented as the final ensemble
870 score. Each of the three streams in module 1 (B) is trained to minimize the error
871 between the predicted "Score" and the score registered in the STITCH database,
872 which contains millions of known and estimated CPIs. The higher the confidence
873 score (closer to 1), the more confident LIGHTHOUSE is that there is some
874 relation between the compound and the protein; conversely, the lower the
875 confidence score (closer to 0), the more confident LIGHTHOUSE is that there is
876 no such relation. Each of the three streams in module 2 (C) is trained to predict
877 scaled IC_{50} values with the use of BindingDB data. For instance, an interaction
878 score of 4 means that, if the compound has inhibitory activity, the IC_{50} would be
879 ~100 μM, whereas an interaction score of 9 means that, if the compound has
880 inhibitory activity, the IC_{50} would be ~1 nM. Note that module 2 only works if the
881 compound and protein interact, so this module is auxiliary to module 1. See also
882 Supplementary Figs. 1,2, and Supplementary Tables 1, 2, 3.
883

884 **Figure 2. *In silico* verification of LIGHTHOUSE**

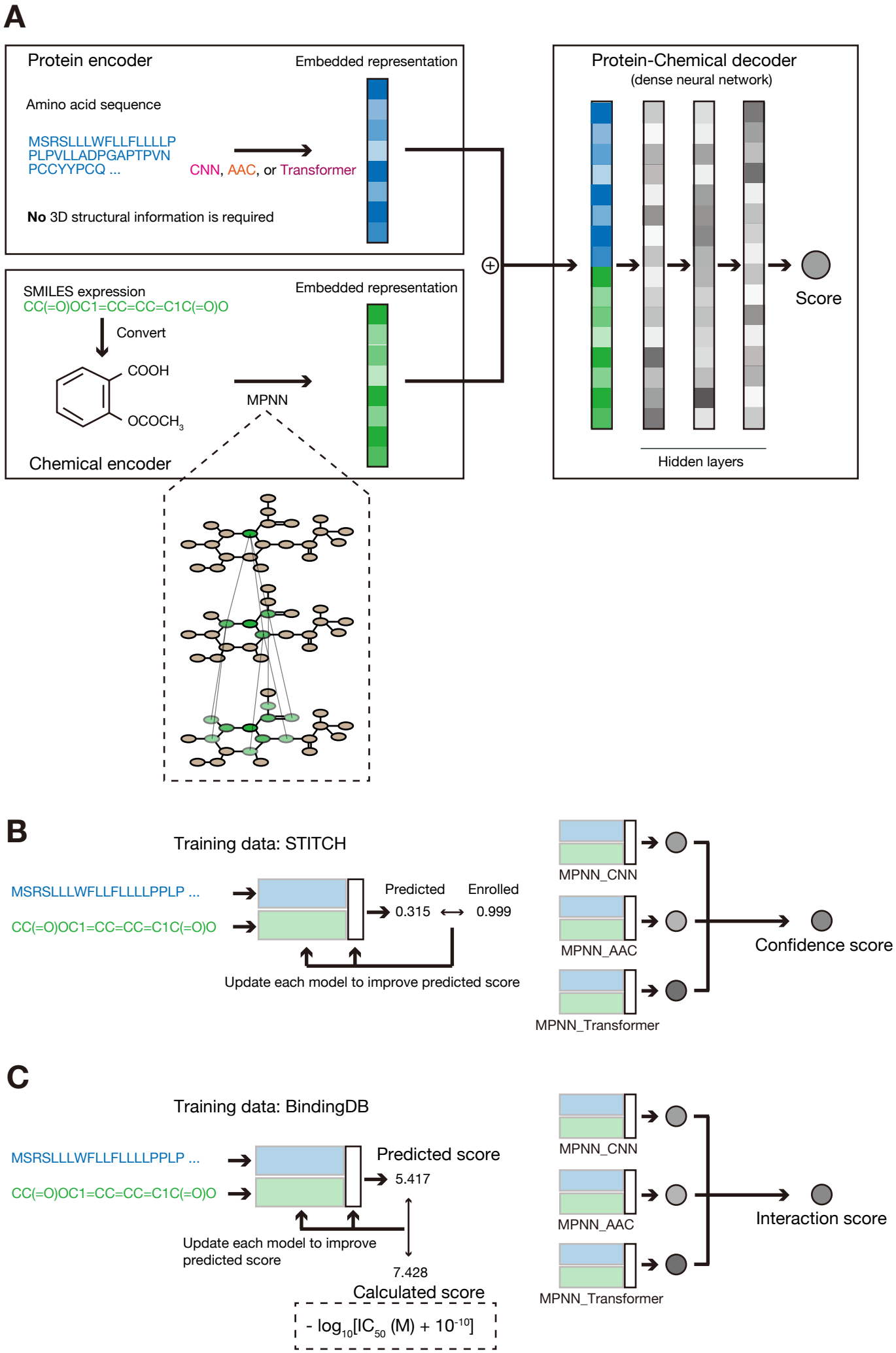
885 (A) For investigation of whether LIGHTHOUSE is able to enrich for compounds
886 with known targets, two data sets were generated from STITCH: a "Positive"
887 data set consisting of CPIs with high scores (>0.9), and a "Negative" data set
888 consisting of the same CPIs but with the amino acid sequences of the proteins
889 reversed (for example, MTSAVM to MVASTM). Proteins in the "Negative" data
890 set would not be expected to interact with the corresponding compounds.

891 LIGHTHOUSE tended to yield higher confidence scores for CPIs in the “Positive”
892 data set, with the exception of the rightmost peak for the “Negative” data set,
893 presumably because these chemicals (such as ATP) are well known and
894 frequently mentioned in the PubMed literature. (B) ROC curve showing that
895 LIGHTHOUSE was able to distinguish the “Positive” and “Negative” data sets.
896 (C–F) Known CPIs and their confidence scores predicted by LIGHTHOUSE. (C)
897 Epinephrine and α-adrenergic (ADRA) and β-adrenergic (ADRB) receptors. (D)
898 Fexofenadine and the histamine receptor HRH1, and famotidine and the
899 histamine receptor HRH2. (E) The macrocyclic drug rapamycin and MTOR. (F)
900 The peptide drugs bortezomib, leuprorelin, and semaglutide and their targets
901 PSMB1, GNRHR, and GLP1R, respectively. These peptide drugs were
902 converted to numerical vectors with the use of SMILES expression and MPNN.
903 (G) Application of LIGHTHOUSE to five drugs approved by the FDA in 2020 that
904 were not included in the training data set (published in 2016). FNTA, protein
905 farnesyltransferase/geranylgeranyltransferase type–1 subunit α; COMT,
906 catechol O-methyltransferase; S1PR1, sphingosine 1-phosphate receptor 1;
907 DRD2, D2 dopamine receptor. See also Supplementary Fig. 3.
908
909 **Figure 3. Discovery of lead compounds for treatment of cancer**
910 (A) Scheme for PPAT inhibitor discovery. The amino acid sequence of PPAT
911 (517 residues) and the SMILE representation for each chemical were entered
912 into the MPNN_CNN model. If the predicted score was >0.5, the compound was
913 entered into MPNN_AAC, and if the new predicted score was >0.5, the
914 compound was entered into MPNN_Transformer. The harmonic mean of the
915 three scores was then computed to obtain the confidence score.
916 (B) Almost 1 billion compounds in the ZINC database were processed as in (A).
917 The first filter (MPNN_CNN score > 0.5) and subsequent two filters (MPNN_AAC
918 score > 0.5, MPNN_Transformer score > 0.5) greatly reduced the initial chemical
919 space (to 0.0356%). The interaction scores for these selected candidates were
920 then also calculated.
921 (C) A 2D map of the 333,290 selected candidates from (B) is shown on the left.
922 Ideal candidates would be expected to have high confidence and interaction
923 scores and would be plotted in the upper right corner of the map. Indeed, some
924 well-known drug-target pairs meet this criterion, as shown on the right, with
925 compounds represented by the blue circles in the shaded area potentially

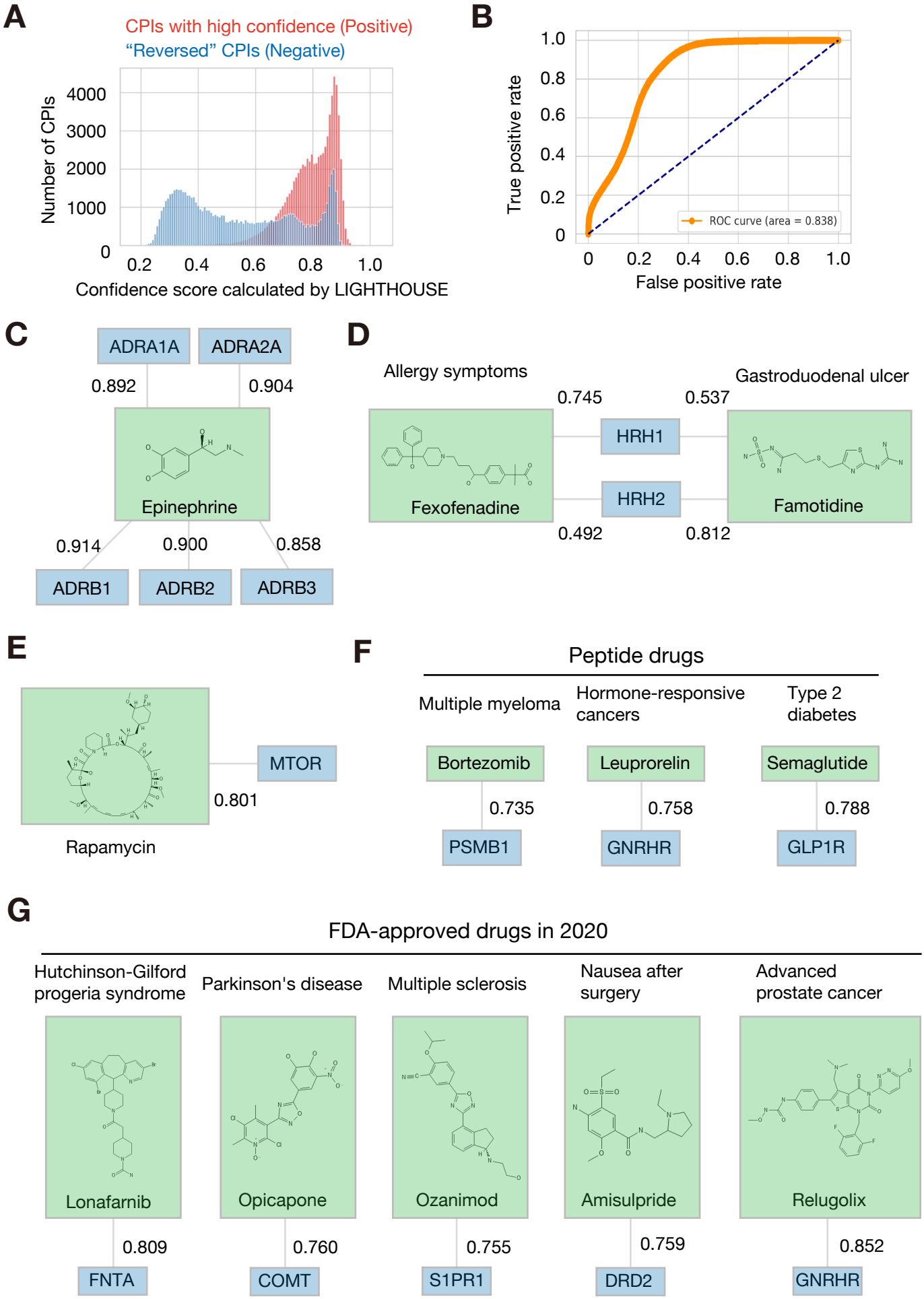
926 possessing inhibitory activity for PPAT. ABL1, ABL proto-oncogene 1; COX1,
927 cyclooxygenase 1; HMG-CoA, 3-hydroxy-3-methylglutaryl–coenzyme A.
928 (D) The top hit compound, ZINC8551105 (riboflavin 5′-monophosphate), is
929 shown together with its confidence score and estimated IC_{50} value.
930 (E) *In vitro* PPAT activity assay performed in the presence of various
931 concentrations (1 nM, 5 nM, 10 nM, 50 nM, 100 nM, 500 nM, 1 μM, 5 μM, 10 μM,
932 50 μM, and 100 μM) of riboflavin 5′-monophosphate, with the determined IC_{50}
933 value being within the range predicted by LIGHTHOUSE. Data are shown for
934 four biological replicates. See also Supplementary Fig. 5.
935
936 **Figure 4. LIGHTHOUSE is applicable to prokaryotic proteins**
937 (A) LIGHTHOUSE predicted that pyridoxal 5′-phosphate would associate with
938 several penicillin binding proteins (PBPs) of *E. coli* (strain K12). MrdA and FtsL
939 are peptidoglycan D,D-transpeptidases, whereas DacA is a D-alanyl-D-alanine
940 carboxypeptidase.
941 (B) *Escherichia coli* strain JM109 was cultured in 2xYT medium supplemented
942 with various concentrations of pyridoxal 5′-phosphate, and optical density at 600
943 nm (OD_{600}) of the culture was monitored. Data are means ± SD for three
944 independent experiments.
945 (C) The JM109 strain of *E. coli* was transformed with the pBlueScript II SK+
946 plasmid, which contains an ampicillin resistance gene as a selection marker, and
947 the cells were plated on LB agar plates containing ampicillin in the absence or
948 presence of pyridoxal 5′-phosphate (3 mg/ml) and were incubated overnight.
949 The pH of pyridoxal 5′-phosphate was adjusted to 7.0 in order to avoid potential
950 nonspecific toxicity.
951
952 **Figure 5. LIGHTHOUSE directs optimization of lead compounds**
953 (A) Prediction of the potency of DPP-4 inhibitor derivatives by LIGHTHOUSE.
954 The predicted interaction scores are compared with the reported IC_{50} values.
955 (B) LIGHTHOUSE accurately predicts that riboflavin is a less potent PPAT
956 inhibitor than is riboflavin 5′-monophosphate.
957 (C) The interaction scores calculated by LIGHTHOUSE for both wild-type (WT)
958 and T315I mutant forms of ABL1 accurately predict that the point mutation
959 reduces the effectiveness of the leukemia drug imatinib.
960
961 **Figure 6. LIGHTHOUSE uncovers potential target proteins for given drugs**

962 (A) Identification of statin targets by LIGHTHOUSE. LIGHTHOUSE was applied
963 to calculate confidence scores for all human protein-coding genes in the UniProt
964 database and fluvastatin, atorvastatin, and cerivastatin. The harmonic mean of
965 these confidence scores (FluvastatinScore, AtorvastatinScore, and
966 CerivastatinScore) was calculated as an affinity score for statins. Sorting on the
967 basis of this affinity score yielded a list of potential statin target proteins.
968 HMGCR (HMG-CoA reductase), a known key target of statins, was ranked 136th
969 with a score of 0.790. The top 500 identified genes were then subjected to
970 enrichment analysis. LDLR, low-density lipoprotein receptor; APOE,
971 apolipoprotein E; SCD, stearoyl-CoA desaturase; STAT3, signal transducer and
972 activator of transcription 3.
973 (B) Distribution of the harmonic mean of the AtorvastatinScore,
974 CerivastatinScore, and FluvastatinScore (affinity score for statins).
975 (C) KEGG pathway enrichment analysis for the top 500 potential statin targets
976 identified by LIGHTHOUSE. Minus $\log_{10}$-transformed q values are shown.
977 (D) Confidence scores for representative predictions by LIGHTHOUSE of the
978 association of statins with cancer-related proteins.
979 (E) Enrichment analysis for expression sites of the top 500 potential statin
980 targets. Minus $\log_{10}$-transformed q values are shown. See also Supplementary
981 Table 5.
982
983 **Figure 7. LIGHTHOUSE-based drug repurposing for COVID-19**
984 (A) Enrichment analysis of the top 500 potential statin targets identified in Fig. 6
985 for COVID-19–associated gene sets. Minus $\log_{10}$-transformed q values are
986 shown.
987 (B) Prediction by LIGHTHOUSE of ethoxzolamide as a potential therapeutic for
988 SARS-CoV-2 infection on the basis of its confidence and interaction scores for
989 ACE2.
990 (C) Vero-TMPRSS2 cells were infected with wild-type SARS-CoV-2 at an MOI of
991 0.0001, cultured for 64 h in the presence of the indicated concentrations of
992 ethoxzolamide, and subjected to immunocytofluorescence analysis with
993 antibodies to SARS-CoV-2 N protein (green). Nuclei were stained with Hoechst
994 33342 (blue). Scale bars, 100 μm.
995 (D and E) Vero-TMPRSS2 cells challenged with wild-type (WK-521) or delta
996 strains of SARS-CoV-2, respectively, were cultured in the presence of various
997 concentrations of ethoxzolamide for 3 days and then subjected to the MTT assay

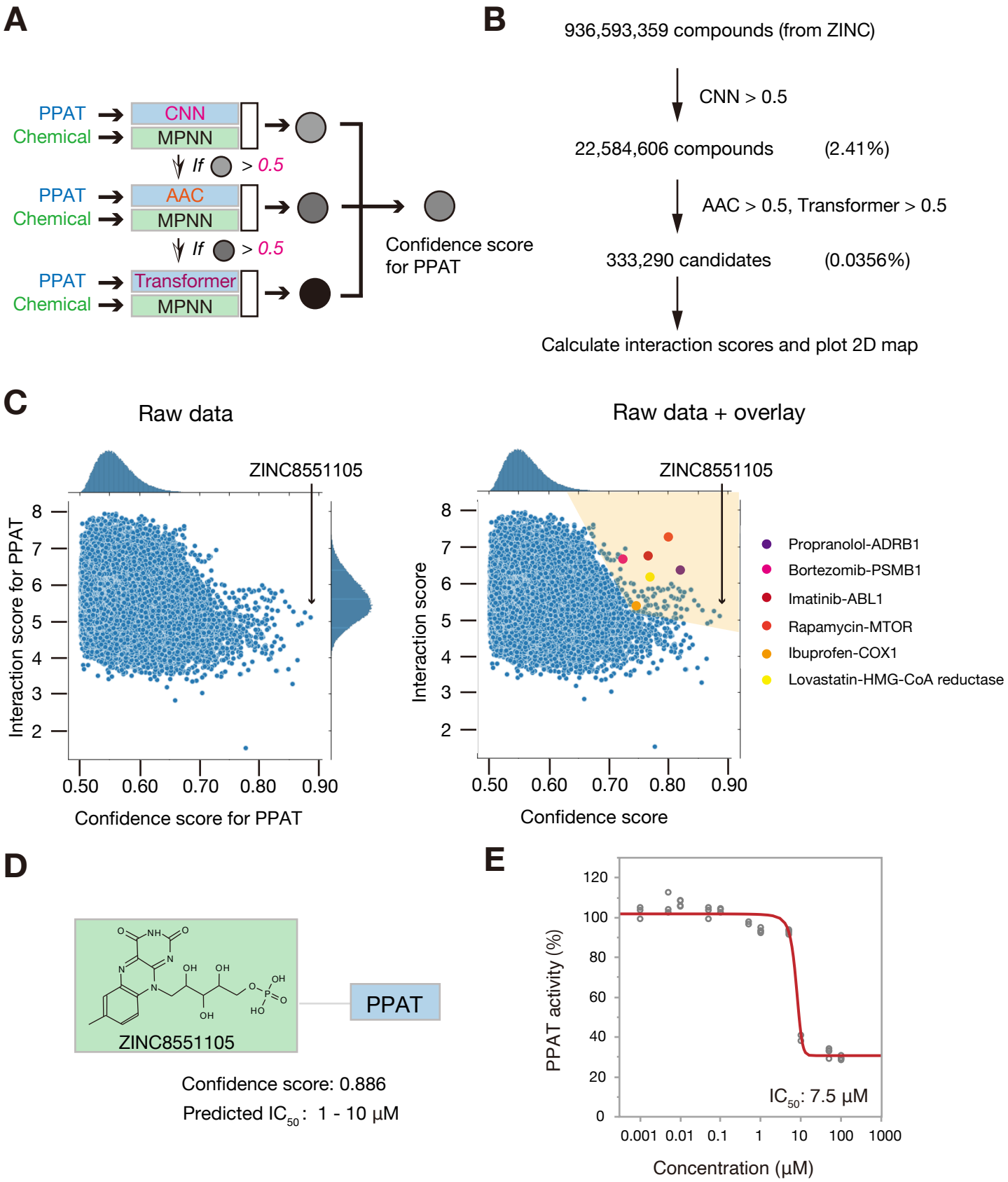
998 of cell viability. Nonchallenged cells were examined as a control. Data are
999 means ± SD for three independent experiments each performed in duplicate.
1000 (F and G) Effect of ethoxzolamide on the SARS-CoV-2 load in culture
1001 supernatants of Vero-TMPRSS2 cells challenged with wild-type (WK-521) or
1002 delta strains of the virus, respectively. Data are from four independent
1003 experiments, with the graph line connecting mean values. $TCID_{50}$, median tissue
1004 culture infectious dose; N.D., not detected. See also Supplementary Figs. 6, 7,
1005 Supplementary Table 6.



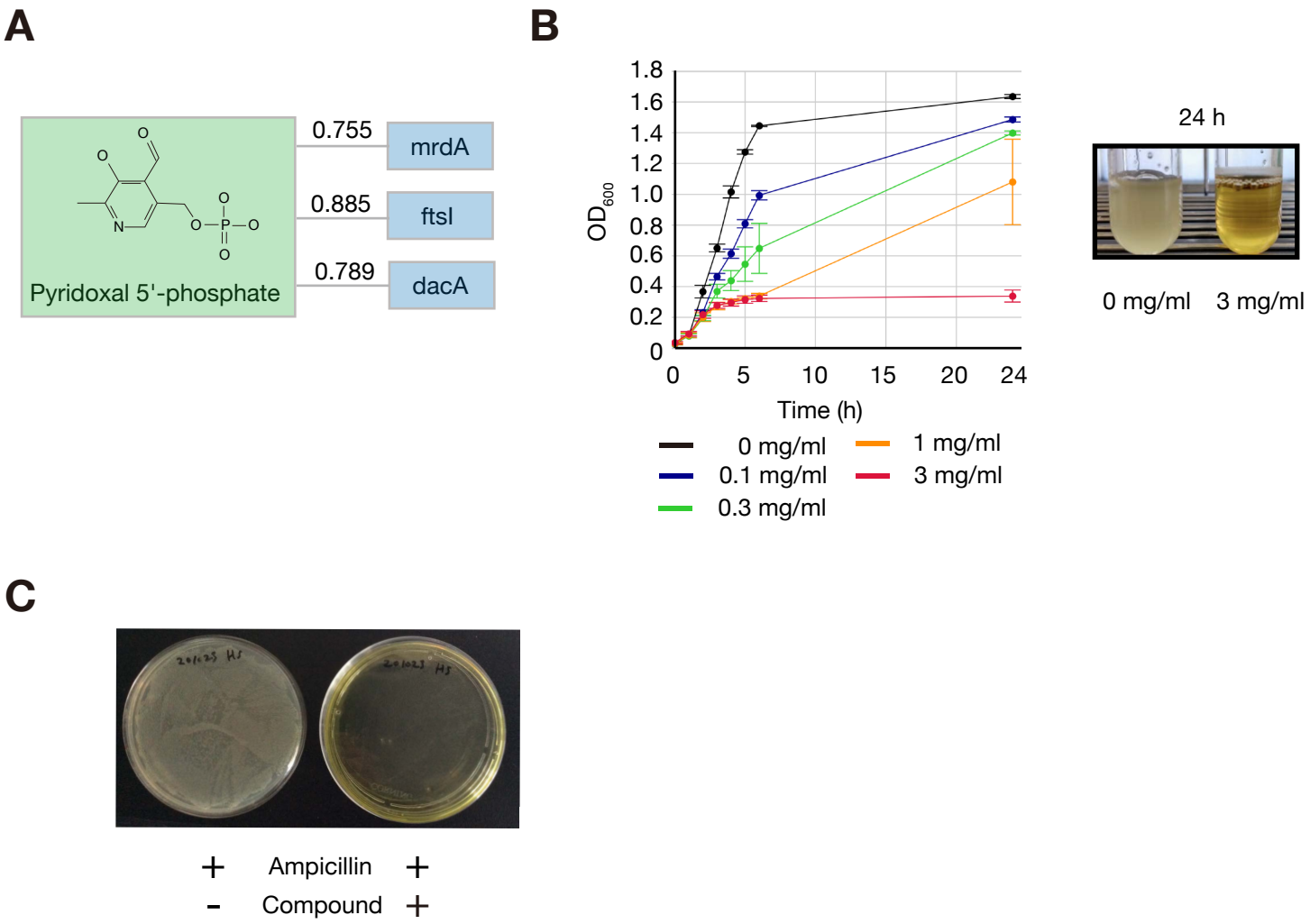
Shimizu et al., Figure 1



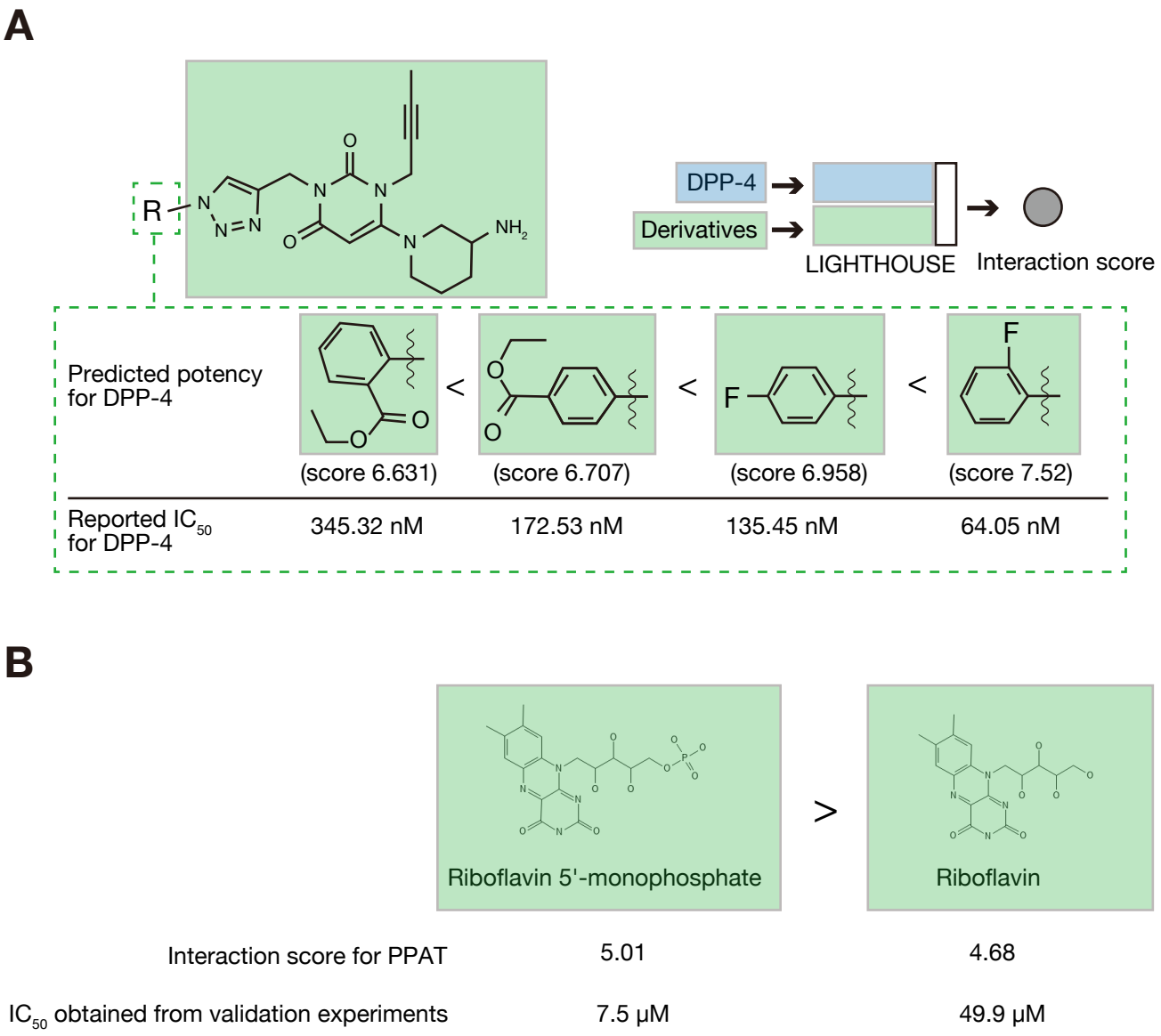
Shimizu et al., Figure 2



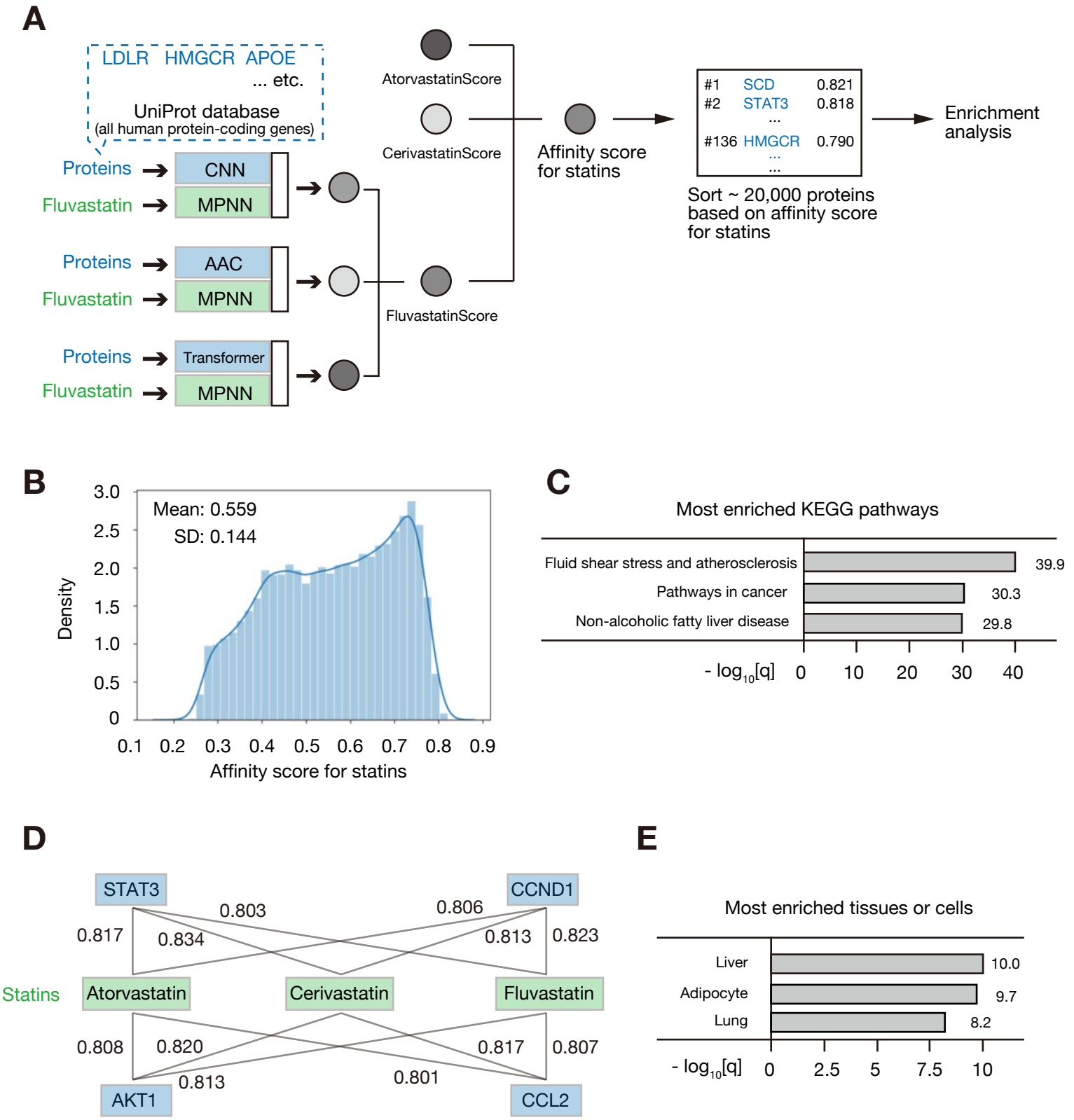
Shimizu et al., Figure 3



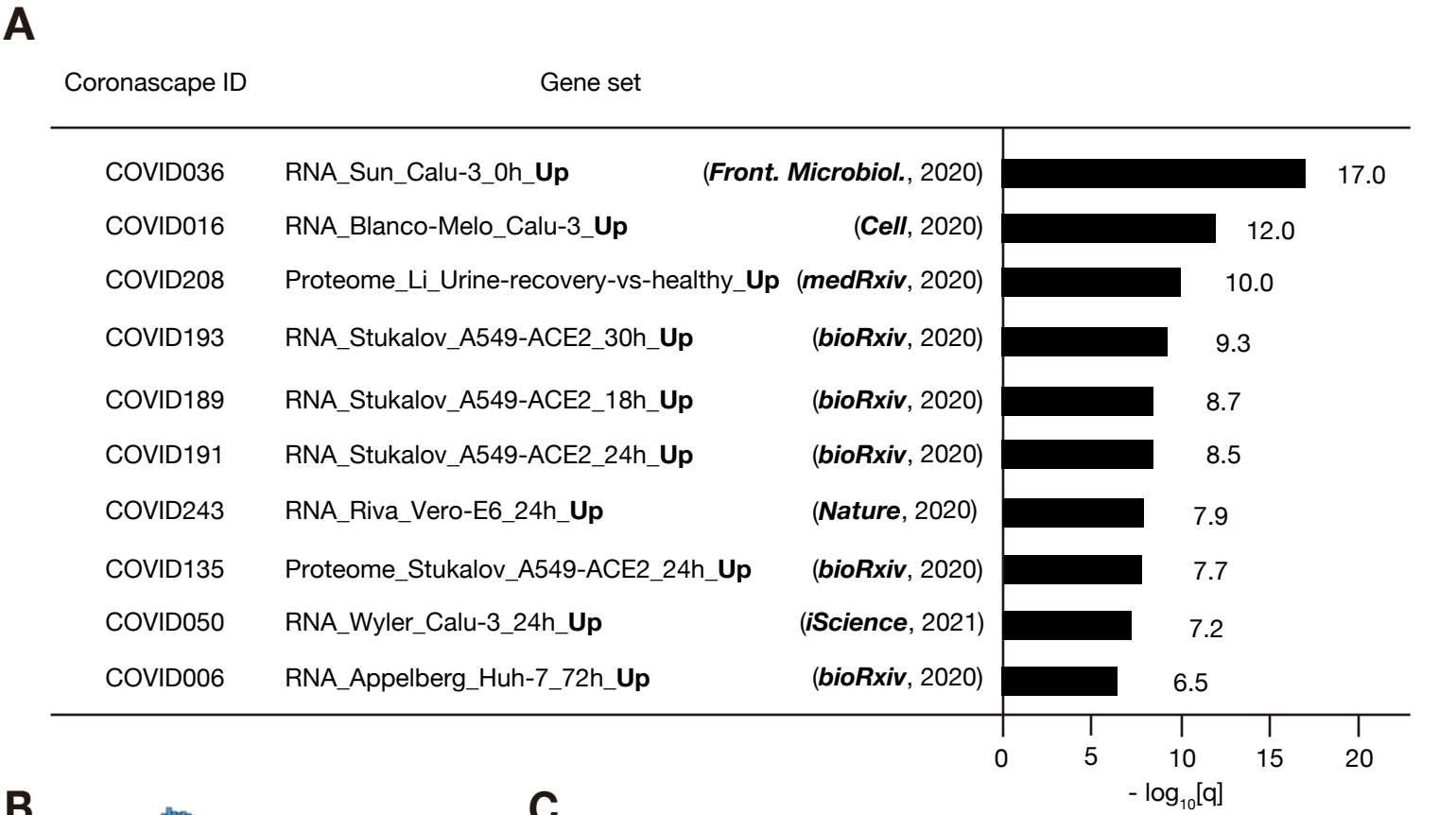
Shimizu et al., Figure 4

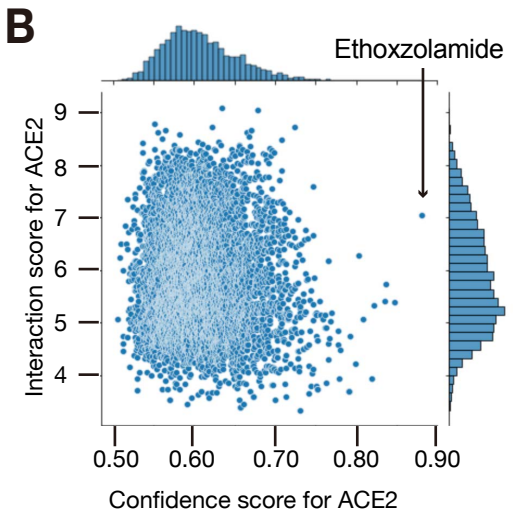


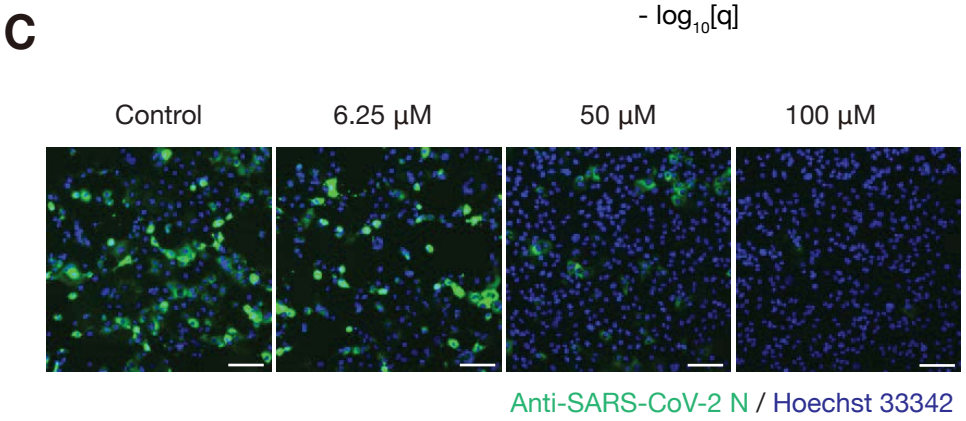
Shimizu et al., Figure 5

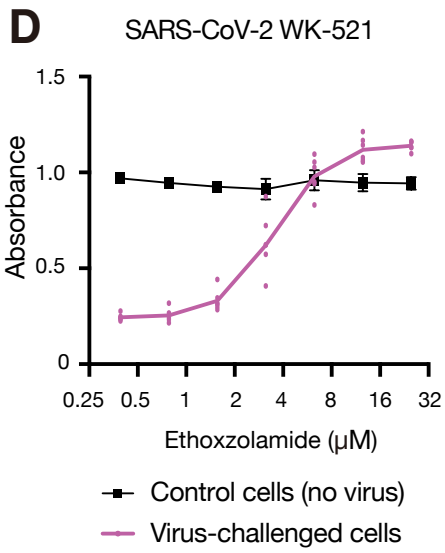


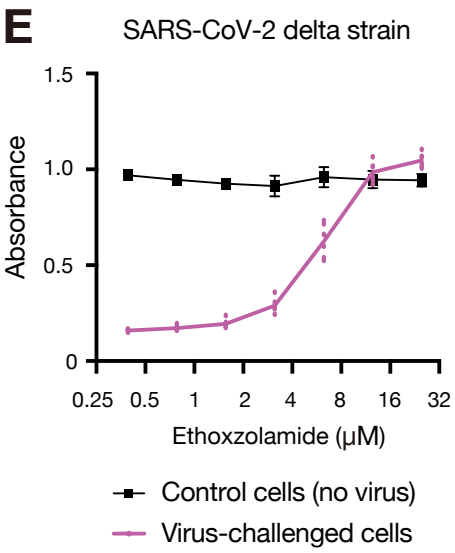
Shimizu et al., Figure 6

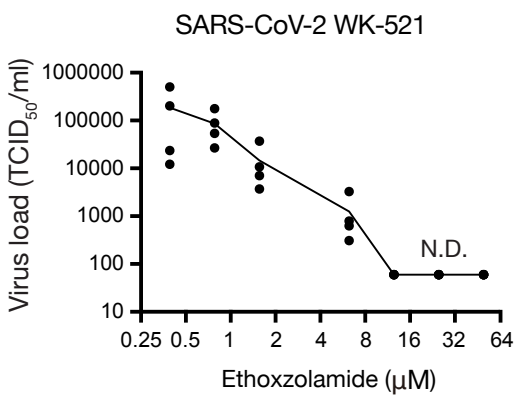


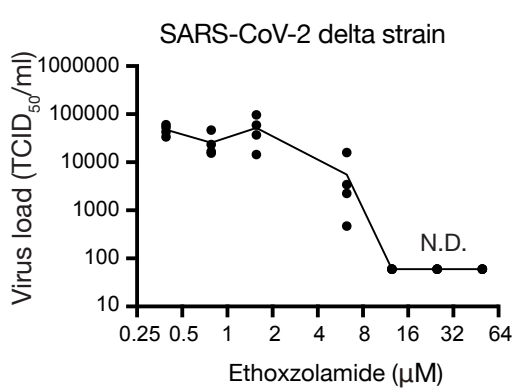












Shimizu et al., Figure 7