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.)

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 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 (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[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 (Figure S1A).
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 (Figure 1B).
132 When we trained the three models (CNN, AAC, and Transformer for protein
133 encoders) separately, the mean squared error (MSE) for the validation data was
134 gradually decreased, and the area under the receiver operating characteristic
135 curve (AUROC) was also improved (Figures S1B–S1G). 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 (Table S2). These scores are equivalent to or
141 better than those of cutting-edge 3D docking simulations.[15–17] It is of note that
142 our AI models can be applied to proteins for which 3D structural information is
143 not available. We took the harmonic mean of the three scores to define the
144 confidence score (Figure 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 (Figure S2A). The same architecture was
149 adopted to train the AI models to predict scaled IC_{50} values (Figure 1C), yielding
150 an interaction score, and we confirmed that the models adequately learned how
151 to predict IC_{50} from amino acid sequence–chemical pairs (Figures S2B–S2G).
152 Finally, we assessed the performance of the models with undisclosed test data,
153 finding that they performed well in predicting IC_{50} (Table S3).
154
155 ***In silico* verification of LIGHTHOUSE**
156 We next evaluated the performance of LIGHTHOUSE in terms of its ability to
157 predict known CPIs. We generated two data sets from STITCH test data: a
158 “Positive” data set consisting of reliable CPIs (STITCH confidence score of >0.9),
159 and a “Negative” data set consisting of CPIs with the lowest confidence scores
160 (<0.2). Calculation by LIGHTHOUSE of the confidence scores for both data sets
161 revealed that those for the "Positive" data set were heavily skewed toward 1,

162 whereas those for the “Negative” data set were almost all <0.5 (Figure 2A).
163 Receiver operating characteristic (ROC) curve analysis showed that the two
164 data sets could be distinguished on the basis of their LIGHTHOUSE confidence
165 scores (Figure 2B). Given that the STITCH database used for the training of
166 LIGHTHOUSE relies not only on experimental CPI data but also on
167 co-appearance of chemicals and proteins in the literature, some well-studied
168 molecules, such as ATP, have high values even in the “Negative" data set.
169 Despite the presence of such false positives, LIGHTHOUSE proved to be
170 effective in predicting the degree of association between protein-chemical pairs
171 solely on the basis of protein primary structure. Removal of molecules smaller
172 than a molecular mass of 200 or 350 Da did not affect the performance of
173 LIGHTHOUSE (Figures S3A–S3D). Furthermore, we also modeled IC_{50} data
174 from BindingDB, which is derived from actual bioassays. We found that the
175 predicted value and observed value correlate well in the BindingDB test data set
176 (Figure S3E).
177 We next validated the effectiveness of LIGHTHOUSE for well-studied
178 compound-protein pairs. LIGHTHOUSE yielded high confidence scores for
179 adrenergic receptors (α1, α2, β1, β2, and β3) and epinephrine (Figure 2C).
180 Histamine receptors are classified into four subtypes,[19] with HRH1 and HRH2
181 being targets of antiallergy and antiulcer drugs, respectively. LIGHTHOUSE
182 predicted that the HRH1 antagonist fexofenadine would associate to a greater
183 extent with HRH1 than with HRH2, whereas the HRH2 inhibitor famotidine would
184 associate to a greater extent with HRH2 than with HRH1 (Figure 2D). These
185 results suggested that LIGHTHOUSE is able to accurately discriminate receptor
186 subtype–level differences solely on the basis of amino acid sequences.
187 LIGHTHOUSE also proved informative both for macrocyclic chemicals
188 such as rapamycin, yielding a high confidence score for this drug and
189 mechanistic target of rapamycin (MTOR) (Figure 2E), as well as for peptide
190 drugs such as bortezomib (used for treatment of multiple myeloma), leuprorelin
191 (hormone-responsive cancers), and semaglutide (type 2 diabetes) (Figure 2F),
192 yielding high confidence scores for these drugs and their known targets:
193 proteasome subunit PSMB1,[20] gonadotropin-releasing hormone receptor
194 (GNRHR),[21] and glucagon-like peptide–1 (GLP-1) receptor (GLP1R),[22]
195 respectively. Given the rapidly growing demand for peptide drugs,[23]
196 LIGHTHOUSE will prove useful for the development of novel peptide
197 therapeutics for a variety of promising targets.

198 We also applied LIGHTHOUSE to five drugs that were approved by the
199 U.S. Food and Drug Administration (FDA) in 2020 but which had not yet been
200 registered in the STITCH database. LIGHTHOUSE successfully predicted the
201 association between these new drugs and their target proteins (Figure 2G),
202 indicating the expandability of LIGHTHOUSE to a much larger exploration space
203 than that encompassed by STITCH. This series of findings thus demonstrated
204 the ability of LIGHTHOUSE to discover new drugs for a broad spectrum of
205 diseases.
206
207 **LIGHTHOUSE discovers an inhibitor of PPAT, a key metabolic enzyme for**
208 **cancer treatment**
209 We investigated whether LIGHTHOUSE can identify compounds for novel and
210 potentially important therapeutic targets. As such a target, we chose
211 phosphoribosyl pyrophosphate amidotransferase (PPAT), a rate-limiting enzyme
212 in the *de novo* nucleotide synthesis pathway, given that its expression is most
213 correlated among all metabolic enzymes with poor prognosis in various human
214 cancers and that its depletion markedly inhibits tumor growth[24]. Although no
215 PPAT inhibitor has been developed and the 3D structure of the protein has not
216 been solved, we attempted to discover an inhibitor for PPAT by LIGHTHOUSE
217 solely on the basis of its amino acid sequence. We virtually screened ~10^9
218 commercially available compounds in the ZINC database[25] (Figure S4). To
219 reduce the calculation time (the MPNN_CNN model took 3 months to screen
220 ~10^9 compounds in the Google Colab Pro environment; Ubuntu 18.04.5 LTS,
221 Intel Xeon CPU @ 2.30 GHz, Memory 25 GB, GPU Tesla P100 16 GB), we
222 adopted a step-by-step application of LIGHTHOUSE (Figure 3A). The
223 MPNN_CNN model excluded most of the chemicals unrelated to PPAT, with
224 only 2.41% of the starting compounds having a score of >0.5 in this initial
225 screening (Figure 3B). The selected compounds were then processed by the
226 MPNN_AAC and MPNN_Transformer models, which reduced the number of
227 candidate chemicals to 0.0356% of the initial compounds. We also calculated
228 interaction scores by LIGHTHOUSE and visualized them in a 2D plot (Figure 3C,
229 left). The best candidates would be expected to have high confidence and
230 interaction scores, appearing in the upper right corner of the plot. Indeed, this
231 criterion was met by several well-known drug-target combinations (Figure 3C,
232 right).

233 Among the >333,000 final compounds, the candidate PPAT inhibitor
234 with the highest confidence score was ZINC8551105 (riboflavin
235 5′-monophosphate), with a predicted IC_{50} of 1 to 10 μM (Figure 3D). We
236 performed a biochemical assay to test this prediction and found that riboflavin
237 5′-monophosphate indeed markedly inhibited PPAT activity with an actual IC_{50} of
238 7.5 μM (Figure 3E). This compound, discovered by LIGHTHOUSE solely on the
239 basis of the PPAT amino acid sequence, is thus a potential lead compound for
240 the development of new therapeutics targeted to a variety of cancers. It is also of
241 note that we tested only this compound, so other top candidates might also
242 inhibit PPAT activity.
243
244 **LIGHTHOUSE identifies an inhibitor of drug-resistant bacterial growth**
245 Bacterial infections pose a clinical problem worldwide, especially in developing
246 countries, and the emergence of drug-resistant bacterial strains as a result of the
247 overuse of antibiotics has exacerbated this problem. β-Lactamase enzymes
248 produced by antibiotic-resistant bacteria[26] target the β-lactam ring of antibiotics
249 of the penicillin family. We therefore applied LIGHTHOUSE to search for
250 antibiotics not dependent on β-lactam structure.
251 LIGHTHOUSE predicted that pyridoxal 5'-phosphate might associate
252 with penicillin binding proteins such as PBP2 (*mrdA*), PBP3 (*ftsI*), and PBP5
253 (*dacA*), all of which are essential for cell wall synthesis in *Escherichia coli*[27]
254 (Figure 4A). This compound indeed suppressed the growth of wild-type *E. coli*
255 strain JM109 in a concentration-dependent manner (Figure 4B). Importantly,
256 pyridoxal 5'-phosphate also markedly inhibited the growth of an
257 ampicillin-resistant *E. coli* transformant that produces β-lactamase (Figure 4C).
258 These results thus suggested that, even though it was trained with human
259 proteins, LIGHTHOUSE can also be applied to nonhuman (even bacterial)
260 proteins.
261
262 **LIGHTHOUSE informs optimization of lead compounds**
263 Diabetes mellitus is also a serious public health concern, with the number of
264 affected individuals expected to increase markedly in the coming decades.[28]
265 Dipeptidyl peptidase–4 (DPP-4) cleaves and inactivates the incretin hormones
266 GLP-1 and glucose-dependent insulinotropic polypeptide (GIP), and DPP-4
267 inhibitors are a new class of antidiabetes drug.[29] Given that LIGHTHOUSE also
268 predicts interaction scores, we examined whether it might also contribute to the

269 optimization step of drug development. Indeed, LIGHTHOUSE accurately
270 predicted the rank order of potency for several recently identified DPP-4 inhibitor
271 derivatives[30] (Figure 5A). Furthermore, LIGHTHOUSE predicted that removal of
272 the phosphate group would reduce the inhibitory potency of riboflavin
273 5′-monophosphate for PPAT (Figure 3E), and this prediction was confirmed
274 correct by the finding that the IC_{50} value was increased from 7.5 to 49.9 μM
275 (Figure 5B). These data suggested the possibility that LIGHTHOUSE is capable
276 of predicting activity cliffs.
277 LIGHTHOUSE is also able to estimate the effect of point mutations on
278 CPIs. For example, the T315I mutation of ABL1 in leukemia cells reduces the
279 efficacy of imatinib,[31] and LIGHTHOUSE accurately predicted the effect of this
280 mutation (Figure 5C). LIGHTHOUSE is able to provide such insight from only
281 wild-type amino acid sequences, given the lack of variant information in the
282 original training data set. Our results suggest that LIGHTHOUSE is able to
283 predict the effects of small changes in protein or chemical structure, and that this
284 will be the case even if such variants do not exist in nature.
285
286 **LIGHTHOUSE identifies potential on- and off-targets of given compounds**
287 Opposite to the mode of drug discovery for a given protein, LIGHTHOUSE might
288 also be able to identify proteins as potential on- or off-targets for a given
289 compound. To verify this notion, we examined statins, which are HMG-CoA
290 reductase inhibitors widely administered for the treatment of hyperlipidemia.
291 Epidemiological studies have shown that statins not only lower cholesterol,
292 however, but also have effects on cancer, although the target molecules for
293 these effects have remained unclear.[32] We therefore applied LIGHTHOUSE to
294 three representative statins (atorvastatin, cerivastatin, and fluvastatin) and
295 computed confidence scores for all human protein-coding genes (Figures 6A
296 and 6B, Table S4). We then sorted the genes on the basis of these confidence
297 scores and performed Kyoto Encyclopedia of Genes and Genomes (KEGG)
298 pathway enrichment analysis for the top 500 potential statin targets. In addition
299 to lipid-related pathways such as atherosclerosis and fatty liver, “pathways in
300 cancer” was one of the most enriched KEGG pathways (Figure 6C), consistent
301 with previous findings.[33–36] Potential targets of statins for cancer treatment
302 identified by LIGHTHOUSE included STAT3, CCND1, AKT1, and CCL2 (Figure
303 6D).

304 Given that side effects of drugs often manifest in organs that express
305 target proteins, we hypothesized that LIGHTHOUSE might be able to identify
306 which organs are at risk of damage from a given drug. We performed another
307 enrichment analysis for the same top 500 potential statin target genes to
308 determine which organs or cell types preferentially express these genes. The top
309 three candidates were the liver, adipocytes, and lung (Figure 6E), consistent
310 with the liver being the primary site of statin metabolism and interstitial
311 pneumonia being one of the most severe side effects of statins.[37] Prediction of
312 potential target proteins for a given drug by LIGHTHOUSE will thus provide
313 insight into which organs warrant close monitoring by physicians during
314 treatment with the drug, especially in first-in-human clinical trials.
315
316 **LIGHTHOUSE identifies novel potential therapeutics for COVID-19**
317 SARS-CoV-2 emerged at the end of 2019 and has caused a pandemic of
318 infectious pulmonary disease, COVID-19.[38] We noticed that genes whose
319 expression is up-regulated after SARS-CoV-2 infection[39–41] were enriched in the
320 list of potential statin targets identified by LIGHTHOUSE (Figure 7A). Indeed,
321 previous studies have shown that statins prevent exacerbation of COVID-19.[42,43]
322 With this finding that LIGHTHOUSE is also effective for COVID-19 drug
323 discovery, we applied it to the virtual screening of ~10,000 approved drugs,
324 given that drug repurposing may allow faster delivery of effective agents to
325 patients in need. We calculated scores for angiotensin-converting enzyme 2
326 (ACE2), which is targeted by SARS-CoV-2 for infection of host cells,[44] and the
327 top drug candidate, ethoxzolamide, was selected for validation analysis (Figure
328 7B). Immunocytofluorescence analysis revealed that ethoxzolamide blocks
329 proliferation of SARS-CoV-2 in Vero-TMPRSS2 cells (Figure 7C). Furthermore,
330 ethoxzolamide was effective against not only the wild-type virus but also the
331 alpha, beta, gamma, and delta variants. It thus rescued virus-challenged cells in
332 a concentration-dependent manner without affecting noninfected cells (median
333 cytotoxicity concentration > 50 μM) (Figures 7D and 7E, Figure S5, Table S5),
334 and it reduced the virus load present in the culture supernatant of the cells
335 (Figures 7F and 7G, Figure S6). Ethoxzolamide is approved for the treatment of
336 seizures and glaucoma,[45,46] and its pharmacodynamics are therefore known. It is
337 therefore immediately available for repurposing for the treatment of patients with
338 COVID-19, with its further optimization having the potential to save many lives.
339 We tested another 11 drugs and found another 2 compounds that also inhibited

340 SARS-CoV-2 infection (Table S6). This high hit rate of 25% (3 hits in 12
341 compounds) shows the potential power of LIGHTHOUSE for drug repositioning.
342
343 **LIGHTHOUSE architecture outperforms state-of-the-art methods**
344 We showed that LIGHTHOUSE is capable of finding lead compounds for a
345 variety of diseases as confirmed by biomedical validation experiments. We
346 finally compared the performance of LIGHTHOUSE with that of similar existing
347 methods. To ensure a fair comparison, we used DUD-E data as were used in
348 previous studies.[7,47] In brief, we randomly split DUD-E data (102 target proteins)
349 into training (72 proteins) and test (30 proteins) data and then trained the
350 LIGHTHOUSE architecture with the DUD-E training data for classification of
351 compounds as active or decoy with regard to the protein in question. After this
352 training, we examined LIGHTHOUSE performance with the DUD-E test data
353 (Figure S7A). Of note, we used only amino acid sequences of the proteins for
354 training and evaluation of LIGHTHOUSE, even though DUD-E provides
355 structural data (as PDB files) for proteins. In addition, we used the balanced data
356 set of DUD-E—the training samples comprise 22,886 active (positive) and
357 22,886 decoy (negative) samples—as in a previous study.[7] LIGHTHOUSE
358 yielded an AUROC for the DUD-E test data of 0.956 (Figure S7B), which was
359 higher than the values produced by state-of-the-art methods including
360 3D-CNN,[48] AtomNet,[47] and the method described by Tsubaki et al.[7] (Figure
361 S7C).
362 For further comparison of LIGHTHOUSE with the second best method,
363 we downloaded CPI data for human and *Caenorhabditis elegans* from the
364 Github repository of Tsubaki et al.[7] Both of these data sets were generated
365 previously.[49] Similar to Tsubaki et al., we retrained the LIGHTHOUSE
366 architecture with these training data, and we found that LIGHTHOUSE
367 outperformed this other method on the basis of both AUROC and F1 metrics
368 (Table S7). These bodies of evidence thus show that LIGHTHOUSE is one of
369 the best architectures for drug discovery available to date.
370
371 **DISCUSSION**
372 Although recent advances in biological and medical research have uncovered
373 various proteins as promising therapeutic targets in a variety of diseases, the
374 clinical application of these research findings has been limited because of the
375 difficulty in identifying therapeutic chemicals for these targets in a cost-effective

376 and high-throughput manner. Acquisition of 3D structural data for target proteins
377 has been labor-intensive, and processing of such data requires a huge amount
378 of computer capacity and time, resulting in a delay in the translation of research
379 findings from the laboratory to the clinic. We have now shown that
380 LIGHTHOUSE facilitates the identification, from a vast chemical space, of
381 candidate compounds for given target proteins solely on the basis of the primary
382 structure of these proteins. Furthermore, the AUROC for LIGHTHOUSE is
383 equivalent to or better than that for state-of-the-art 3D docking simulation
384 methods as well as for other AI methodologies.
385 We have applied LIGHTHOUSE to attractive targets for various
386 diseases, including cancer, bacterial infection, metabolic diseases, and
387 COVID-19, with some of the suggested chemicals being determined
388 experimentally to be effective for inhibition of the corresponding targets. The
389 actual targets of identified compounds require further biological investigation,
390 especially in the case of pyridoxal 5'-phosphate and PBPs. As for ethoxzolamide,
391 we attempted to obtain clinical evidence in support of its effectiveness against
392 COVID-19. However, ethoxzolamide is an old drug and is essentially no longer
393 prescribed as a result of the development of more potent agents such as
394 acetazolamide. We were therefore not able to find data to address whether
395 COVID-19 patients taking ethoxzolamide have a better clinical outcome.
396 The hit compounds themselves identified in the present study are not
397 sufficient for immediate clinical use. LIGHTHOUSE can be applied not only for
398 the identification of lead compounds but also for their subsequent optimization,
399 which requires extensive work to identify more potent and specific or less toxic
400 derivatives. One promising method to support such optimization is to apply
401 LIGHTHOUSE and either reinforcement learning[50] (Figure S8) or
402 Metropolis-Hasting (MH) approaches together. Virtual libraries can be generated
403 from identified lead compounds in an intensive manner with the use of
404 sophisticated chemoinformatics algorithms such as RECAP (Retrosynthetic
405 Combinatorial Analysis Procedure).[51] LIGHTHOUSE can score the generated
406 virtual compounds and help to narrow down the candidates with better scores
407 than the original hit compound. Select candidates can then be synthesized in
408 collaboration with organic chemists and their effects tested. Given the recent
409 success of the MH approach in various life science fields,[52,53] LIGHTHOUSE
410 should be able to facilitate the optimization of drug candidates by iterating these
411 steps.

412 The biggest limitation of LIGHTHOUSE is the generation of false
413 positives, which is due in part to the fact that the confidence score provided by
414 STITCH is not based solely on experimental data but also on other factors such
415 as co-occurrence in the literature. This confidence score therefore does not
416 necessarily reflect actual interaction for a given protein-chemical pair, and
417 well-studied molecules are thus prone to score higher than others. On the other
418 hand, low confidence scores do not necessarily mean that the protein and
419 chemical in question do not interact. We modeled this score because current
420 biologically determined CPI data sets contain fewer CPI pairs. In addition, as a
421 result of publication bias and biological experimental conditions, it is sometimes
422 difficult to tell whether a protein and chemical do not interact, whether they did
423 not bind under the specific assay condition, or whether they were not tested.
424 This drawback of LIGHTHOUSE can be mitigated partially by combining the
425 three different models (CNN, AAC, and Transformer). It may also be important to
426 perform a counter–virtual screening to determine whether an identified small
427 molecule reacts specifically with the target protein or whether it scores highly
428 with many proteins. Such an approach has the potential to reduce the number of
429 false positives and provide more accurate guidance. Furthermore, we also
430 modeled IC_{50} data from BindingDB, which is derived from actual bioassays. We
431 found that the predicted value and observed value correlate well in the
432 BindingDB test data set (Figure S3E). By combining the confidence score (from
433 STITCH) and the interaction score (from BindingDB) provided by LIGHTHOUSE,
434 we were able to discover a clinically approved drug that was able to block
435 SARS-CoV-2 infection.
436 Another potential limitation is that the performance for STITCH data
437 might be overestimated because we downsampled the original imbalanced data
438 (Table S1). There are several ways to tackle imbalanced data, including
439 downsampling, oversampling, and more complex sampling methods such as
440 SMOTE.[54] Each of these approaches has potential limitations, however.[55,56] In
441 the present study, we adopted downsampling, as in a previous study,[7] so the
442 performance for STITCH data may be overestimated when compared with use
443 of the entire STITCH data.
444 Despite these limitations, LIGHTHOUSE proved to be effective for the
445 identification of lead compounds for all conditions tested. It can theoretically be
446 applied to any protein of any organism, and even to proteins that do not exist
447 naturally. This is an advantage over 3D docking simulation methods, which

448 require prior 3D structural knowledge of the protein of interest. LIGHTHOUSE
449 computes and embeds structural information in numerical vectors, which are
450 then readily retrieved by the subsequent decoding module. Given the
451 accelerating development of protein embedding technologies[57] and graph-based
452 chemoinformatics approaches, LIGHTHOUSE has the potential to be a
453 cornerstone of drug discovery. It is also of note that we split the STITCH data set
454 once, given that a previous study[7] showed it could obtain high performance with
455 a relatively small range of hyperparameter tuning. In-depth hyperparameter
456 tuning with cross-validation may further boost the performance of the model.
457 The next step for more sophisticated AI-dependent drug discovery and
458 its clinical application will be the use of huge-scale and robust protein-chemical
459 binding (and nonbinding) data, with the recent introduction of academic journals
460 that specialize in big data reflecting the growth of this field. Moreover,
461 development of improved approaches to the handling of imbalanced data[58] is
462 another active research field in computer science. These advances in both data
463 and methodology should allow more reliable modeling of physical interactions
464 and further facilitate drug discovery in the coming years.
465 In summary, we have developed LIGHTHOUSE as a means to discover
466 promising lead compounds for any target protein irrespective of its 3D structural
467 information. Furthermore, we have demonstrated the power of LIGHTHOUSE by
468 identifying and validating novel therapeutics for various global health concerns
469 including COVID-19. LIGHTHOUSE will serve as a guide for researchers in all
470 areas of biomedicine, paving the way for a wide range of future applications.
471

472 **EXPERIMENTAL PROCEDURES**

473 **Resource availability**

474 ***Lead contact***

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

478 ***Materials availability***

479 This study did not generate new unique reagents.

480 ***Data and code availability***

481 All data used for training are available from STITCH (http://stitch.embl.de)[14] and
482 BindingDB (https://www.bindingdb.org/bind/index.jsp)[18] web servers. SMILES
483 representations for small-molecule compounds were downloaded from

PubChem (https://pubchem.ncbi.nlm.nih.gov) or ZINC15 (https://zinc15.docking.org).[25] Amino acid sequences were obtained from UniProt (https://www.uniprot.org).[59] For drug repurposing analyses, we used the KEGG-DRUG database (https://www.genome.jp/kegg/drug).[60] All referenced COVID-19 signatures are available at Coronascape (https://metascape.org/COVID).[61]

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.

Generation of a data set for the training phase of LIGHTHOUSE

The compound SMILES strings of the data set were extracted from the PubChem compound database on the basis of compound names and PubChem compound IDs (CIDs). The protein sequences of the data set were extracted from the UniProt protein database on the basis of gene names/RefSeq accession numbers or the UniProt IDs. We downloaded the protein-chemical link data set of *Homo sapiens* (Taxonomy ID 9606) from the STITCH database (version 5.0). Given that the STITCH score is heavily biased toward 0, we separated the data into nine bins on the basis of the score and stratify-extracted the same number of CPIs (140,000 each), yielding 1,260,000 CPIs (Table S1). We then randomly separated these data into training (80%), validation (10%), and test (10%) data sets (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%) data sets (Figure S2A). Given that IC_{50} values differ widely, we scaled the values by log transformation (Eq. 1) and used the transformed values for BindingDB training.

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

LIGHTHOUSE architecture and training

The proposed overall model comprises two encoder networks (for chemicals and proteins) and one decoder network. MPNN is a message passing graph neural network that operates on compound molecular graphs.[11] In brief, MPNN conveys latent information among the atoms and edges. The message passing phase runs for t time steps and is defined in terms of message functions M_t and

vertex update functions U_t. During this phase, hidden states h_v^t (128 dimensions in our model) at each node in the chemical graph are updated with the incoming messages m_v^{t+1} according to the following equations (Eqs. 2 and 3):

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

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

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):

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

where P_i is the LIGHTHOUSE-predicted score for the ith compound-protein pair and Y_i is the true label in the corresponding training data, with a batch size of 128. We trained three architectures (MPNN_CNN, MPNN_AAC, MPNN_Transformer) separately for the STITCH and BindingDB training data with the Adam optimizer and a learning rate of 0.001. For evaluation metrics, we used MSE, concordance index, and Pearson correlation as well as AUROC. For every 10 epochs, we compared the current loss (in the validation data set) with that of 10 epochs ago; if the loss was not decreasing, we terminated the training for that model. As a result of this early termination, we trained MPNN_CNN for 40 epochs, MPNN_AAC for 70 epochs, and MPNN_Transformer for 100 epochs with regard to the confidence score (Figure 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 (Figure S2B–S2G), according to the same guidelines. After the training was completed, we finally evaluated the models with the test data sets, which were kept aside during the training and so had not previously been seen by the models.

Generation of virtual chemical libraries and prediction by LIGHTHOUSE

We prepared nearly 1 billion purchasable substances, which were downloaded from the ZINC database[25] as of 30 July 2020, for virtual PPAT inhibitor screening. For drug repurposing, we obtained approved drugs from the KEGG-DRUG database[60] 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.

PPAT activity assay

587 Sf21 cells were cultured in Sf-900 II SFM (Gibco, Cat# 10902-088)
588 supplemented with 10 μM ferric ammonium citrate. They were transfected with a
589 bacmid encoding human PPAT for 64 h, harvested, washed three times with
590 phosphate-buffered saline, and lysed in a solution containing 150 mM NaCl, 25
591 mM Tris-HCl (pH 7.4), 0.5% Triton X-100, and 5 mM EDTA. The lysate was
592 centrifuged at 10,000 × *g* for 6 min at 4°C, and the resulting supernatant (100
593 ng/ml) was incubated for 4 h at 37°C together with 5 mM glutamine (Gibco, Cat#
594 25030-081), 1 mM phosphoribosyl pyrophosphate (Sigma, Cat# P8296), 10 mM
595 $MgCl_2$, 50 mM Tris-HCl (pH 7.4), and various concentrations of riboflavin
596 5′-monophosphate (Sigma, Cat# F2253-10 mg). Enzyme activity was assessed
597 on the basis of glutamate production as measured with a glutamate assay kit
598 (Abcam, Cat# 138883). The IC_{50} value was estimated from biological
599 quadruplicates with a four-parameter logistic model[63] and with the use of JMP
600 pro 15 software (version 15.1.0).
601
602 **Assay of *E. coli* growth**
603 Portions (20 *μ*l) of *E. coli* strain JM109 (1×10^{10} colony-forming units (CFU)/ml)
604 were cultured for various times in 2 ml of 2xYT liquid medium (BD Difco, Cat#
605 244020) containing various concentrations of pyridoxal 5′-phosphate (pH
606 adjusted to 7.0), after which OD_{600} was measured with a GENESYS 30 visible
607 spectrophotometer (ThermoFisher Scientific, Cat# 840-277000). In addition, the
608 JM109 strain was transformed with 1 μg of the pBlueScript II SK+ plasmid
609 (Invitrogen), which harbors an ampicillin resistance gene as a selection marker,
610 and was then spread on LB agar plates containing ampicillin (100 μg/ml) (Wako,
611 Cat# 012-23303) with or without pyridoxal 5′-phosphate (3 mg/ml) and incubated
612 overnight.
613
614 **Virtual identification of statin targets and enrichment analyses**
615 Three representative statins were fixed as chemical inputs, and all human
616 protein-coding genes in the UniProt database were iteratively changed. The
617 harmonic mean of the three confidence scores was calculated as an affinity
618 score for statins, and the human protein-coding genes were sorted on the basis
619 of this score. The resulting top 500 potential targets were then subjected to
620 enrichment analyses with the use of the Metascape web server.[61]
621
622 **SARS-CoV-2 assays**

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

SUPPLEMENTAL INFORMATION

Supplemental information can be found onine at....

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 from Takeda Science Foundation and Mochida Foundation. We thank T. Sawada and R. Wada for

technical assistance, P.A. Silver and J.C. Way for critical reading of the manuscript, other laboratory members for discussion, and A. Ohta for help with preparation of the manuscript.

AUTHOR CONTRIBUTIONS

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

DECLARATION OF INTERESTS

The authors declare no competing interests.

673 **REFERENCES**

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

708 14. Szklarczyk D., Santos A., Von Mering C., Jensen L.J., Bork P., Kuhn M.
709 STITCH 5: augmenting protein-chemical interaction networks with tissue and
710 affinity data. *Nucleic Acids Res.* 2016; **44**: D380–D384.
711 15. Hsin K.-Y., Matsuoka Y., Asai Y., Kamiyoshi K., Watanabe T., Kawaoka Y.,
712 Kitano H. systemsDock: a web server for network pharmacology-based
713 prediction and analysis. *Nucleic Acids Res.* 2016; **44**: W507–W513.
714 16. Wang S., Jiang J.-H., Li R.-Y., Deng, P. Docking-based virtual screening of
715 TβR1 inhibitors: evaluation of pose prediction and scoring functions. *BMC Chem.*
716 2020; **14**: 52.
717 17. Moussa N., Hassan A., Gharaghani S. Pharmacophore model, docking,
718 QSAR, and molecular dynamics simulation studies of substituted cyclic imides
719 and herbal medicines as COX-2 inhibitors. *Heliyon* 2021; **7**: e06605.
720 18. Gilson M.K., Liu T., Baitaluk M., Nicola G., Hwang L., Chong J. BindingDB in
721 2015: a public database for medicinal chemistry, computational chemistry and
722 systems pharmacology. *Nucleic Acids Res.* 2016; **44**: D1045–D1063.
723 19. Seifert R., Strasser A., Schneider E.H., Neumann. D., Dove S., Buschauer A.
724 Molecular and cellular analysis of human histamine receptor subtypes. *Trends*
725 *Pharmacol. Sci.* 2013; **34**: 33–58.
726 20. Berkers C.R., Verdoes M., Lichtman E., Fiebiger E., Kessler B.M., Anderson
727 K.C., Ploegh H.L., Ovaa H., Galardy P.J. Activity probe for in vivo profiling of the
728 specificity of proteasome inhibitor bortezomib. *Nat. Methods* 2005; **2**: 357–362.
729 21. Borroni R., Blasio A.M.D., Gaffuri B., Santorsola R., Busacca M., Viganò P.,
730 Vignali M. Expression of GnRH receptor gene in human ectopic endometrial
731 cells and inhibition of their proliferation by leuprolide acetate. *Mol. Cell.*
732 *Endocrinol.* 2000; **159**: 37–43.
733 22. Knudsen L.B, Lau J. The discovery and development of liraglutide and
734 semaglutide. *Front. Endocrinol. (Lausanne)* 2019; **10**: 155.
735 23. Muttenthaler M., King G.F., Adams D.J., Alewood P.F. Trends in peptide
736 drug discovery. *Nat. Rev. Drug Discov.* 2021; **20**: 309–325.
737 24. Kodama M, Oshikawa K, Shimizu H, Yoshioka S, Takahashi M., Izumi Y.,
738 Bamba T., Tateishi C., Tomonaga T., Matsumoto M. et al. A shift in glutamine
739 nitrogen metabolism contributes to the malignant progression of cancer. *Nat.*
740 *Commun.* 2020; **11**:1320.
741 25. Sterling T., Irwin J.J. ZINC 15—ligand discovery for everyone. *J. Chem. Inf.*
742 *Model.* 2015; **55**: 2324–2337.

743 26. Tooke C.L., Hinchliffe P., Bragginton E.C., Colenso C.K., Hirvonen V.H.A.,
744 Takebayashi Y., Spencer J. β-Lactamases and β-lactamase inhibitors in the
745 21st century. *J. Mol. Biol.* 2019; **431**: 3472–3500.
746 27. Macheboeuf P., Contreras-Martel C., Job C., Dideberg O., Dessen A.
747 Penicillin binding proteins: key players in bacterial cell cycle and drug resistance
748 processes. *FEMS Microbiol. Rev.* 2006; **30**: 673–691.
749 28. Zheng Y., Ley S.H., Hu F.B. Global aetiology and epidemiology of type 2
750 diabetes mellitus and its complications. *Nat. Rev. Endocrinol.* 2018; **14**: 88–98.
751 29. Deacon C. F. Dipeptidyl peptidase 4 inhibitors in the treatment of type 2
752 diabetes mellitus. *Nat. Rev. Endocrinol.* 2020; **16**: 642–653.
753 30. Li Q., Han L., Zhang B., Zhou J., Zhang H. Synthesis and biological
754 evaluation of triazole based uracil derivatives as novel DPP-4 inhibitors. *Org.*
755 *Biomol. Chem.* 2016; **14**: 9598–9611.
756 31. Jabbour E., Kantarjian H., Jones D., Breeden M., Garcia-Manero G., O'Brien
757 S., Ravandi F., Borthakur G., Cortes J. Characteristics and outcomes of patients
758 with chronic myeloid leukemia and T315I mutation following failure of imatinib
759 mesylate therapy. *Blood* 2008; **112**: 53–55.
760 32. Mei Z., Liang M., Li L., Zhang Y., Wang Q., Yang W. Effects of statins on
761 cancer mortality and progression: a systematic review and meta-analysis of 95
762 cohorts including 1,111,407 individuals. *Int. J. Cancer* 2017; **140**: 1068–1081.
763 33. Matusewicz L., Meissner J., Toporkiewicz M., Sikorski A.F. The effect of
764 statins on cancer cells—review. *Tumour Biol.* 2015; **36**: 4889–4904.
765 34. Ahern T.P., Lash T.L., Damkier P., Christiansen P.M., Cronin-Fenton D.P.
766 Statins and breast cancer prognosis: evidence and opportunities. *Lancet Oncol.*
767 2014; **15**: e461–e468.
768 35. Mullen P.J., Yu R., Longo J., Archer M.C., Penn L.Z. The interplay between
769 cell signalling and the mevalonate pathway in cancer. *Nat. Rev. Cancer* 2016;
770 **16**: 718–731.
771 36. Alfaqih M.A., Allott E.H., Hamilton R.J., Freeman M.R., Freedland S.J. The
772 current evidence on statin use and prostate cancer prevention: Are we there
773 yet? *Nat. Rev. Urol.* 2017; **14**: 107–119.
774 37. Momo K., Takagi A., Miyaji A., Koinuma M. Assessment of statin-induced
775 interstitial pneumonia in patients treated for hyperlipidemia using a health
776 insurance claims database in Japan. *Pulm. Pharmacol. Ther.* 2018; **50**: 88–92.
777 38. Hu B., Guo H., Zhou P., Shi Z.-L. Characteristics of SARS-CoV-2 and
778 COVID-19. *Nat. Rev. Microbiol.* 2021; **19**: 141–154.

39. 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.
40. 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.
41. 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.
42. 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.
43. 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.
44. 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.
45. 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.
46. Ghorai S., Pulya S., Ghosh K., Panda P., Ghosh B., Gayen S. Structure-activity relationship of human carbonic anhydrase-II inhibitors: detailed insight for future development as anti-glaucoma agents. *Bioorg. Chem.* 2020; **95**: 103557.
47. Wallach I., Dzamba M., Heifets A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. *arXiv* 2015, 1510.02855 [cs.LG].
48. Ragoza M., Hochuli J., Idrobo E., Sunseri J. and Koes D. R. Protein-Ligand Scoring with Convolutional Neural Networks. *J Chem Inf Model.* 2017; **57**: 942-957.

814 49. Liu H., Sun J., Guan J., Zheng J., Zhou S. Improving compound–protein
815 interaction prediction by building up highly credible negative samples.
816 *Bioinformatics* 2015; **31**: i221–i229.
817 50. Pereira T., Abbasi M., Ribeiro B., Arrais J.P. Diversity oriented deep
818 reinforcement learning for targeted molecule generation. *J. Cheminform.* 2021;
819 **13**: 21.
820 51. Lewell X.Q., Judd D.B., Watson S.P., Hann M.M. RECAP—retrosynthetic
821 combinatorial analysis procedure: a powerful new technique for identifying
822 privileged molecular fragments with useful applications in combinatorial
823 chemistry. *J. Chem. Inf. Comput. Sci.* 1998; **38**: 511–522.
824 52. Biswas S., Khimulya G., Alley E.C., Esvelt K.M., Church G.M. Low-N protein
825 engineering with data-efficient deep learning. *Nat. Methods* 2021; **18**: 389–396.
826 53. Chen X., Zhou J., Zhang R., Wong A.K., Park C.Y., Theesfeld C.L.,
827 Troyanskaya O.G. Tissue-specific enhancer functional networks for associating
828 distal regulatory regions to disease. *Cell Syst.* 2021; **12**: 353–362.
829 54. Chawla N.V., Bowyer K.W., Hall L. O., Kegelmeyer W.P. SMOTE: Synthetic
830 Minority Over-sampling Technique. *arXiv:1106.1813 [cs.AI]*
831 55. Yu L, Zhou N. Survey of Imbalanced Data Methodologies. *arXiv:2104.02240*
832 *[stat.ML]*
833 56. Kulkarni A, Chong D, Batarseh F.A. Foundations of data imbalance and
834 solutions for a data democracy. *arXiv:2108.00071 [cs.LG]*
835 57. Bepler T., Berger B. Learning the protein language: evolution, structure, and
836 function. *Cell Syst*. 2021; **12**: 654–669.
837 58. Park S, Lim J, Jeon Y, Choi J.Y. Influence-Balanced Loss for Imbalanced
838 Visual Classification. *arXiv:2110.02444 [cs.CV]*
839 59. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021.
840 *Nucleic Acids Res.* 2021; **49**: D480–D489.
841 60. Kanehisa M., Furumichi M., Sato Y., Ishiguro-Watanabe M., Tanabe M.
842 KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* 2021; **49**:
843 D545–D551.
844 61. Zhou Y., Zhou B., Pache L., Chang M., Khodabakhshi A. H., Tanaseichuk O.,
845 Benner C., Chanda S.K. Metascape provides a biologist-oriented resource for
846 the analysis of systems-level datasets. *Nat. Commun.* 2019; **10**: 1523.
847 62. Shimizu H., Nakayama K.I. Artificial intelligence in oncology. *Cancer Sci.*
848 2020; **111**: 1452–1460.

849 63. Pries V., Nöcker C., Khan D., Johnen P., Hong Z., Tripathi A., Keller A.-L.,
850 Fitz M., Perruccio F., Filipuzzi I. Target identification and mechanism of action of
851 picolinamide and benzamide chemotypes with antifungal properties. *Cell Chem.*
852 *Biol.* 2018; **25**: 279–290.
853 64. Sasaki M., Uemura K., Sato A., Toba S., Sanaki T., Maenaka K., Hall W.W.,
854 Orba Y., Sawa H. SARS-CoV-2 variants with mutations at the S1/S2 cleavage
855 site are generated in vitro during propagation in TMPRSS2-deficient cells. *PLoS*
856 *Pathog*. 2021; **17**: e1009233.
857

858 **FIGURE LEGENDS**

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

861 (A) The basic network structure of LIGHTHOUSE consists of encoder and
862 decoder networks. The basic network encodes the amino acid sequence of the
863 protein of interest as numerical vectors by one of three independent methods:
864 CNN, AAC, and Transformer. It also takes the SMILES representation of each
865 small-molecule compound and computes the neural representation with the
866 MPNN algorithm. The network then concatenates the protein and compound
867 representations and calculates a "Score."
868 (B and C) LIGHTHOUSE consists of two modules. Module 1 estimates the
869 association between a given compound-protein pair, and module 2 predicts a
870 scaled IC_{50} value for the pair. In each module, the three different streams of the
871 basic network (MPNN_CNN, MPNN_AAC, and MPNN_Transformer) are used,
872 and the harmonic mean of the three scores is presented as the final ensemble
873 score. Each of the three streams in module 1 (B) is trained to minimize the error
874 between the predicted "Score" and the score registered in the STITCH database,
875 which contains millions of known and estimated CPIs. The higher the confidence
876 score (closer to 1), the more confident LIGHTHOUSE is that there is some
877 relation between the compound and the protein; conversely, the lower the
878 confidence score (closer to 0), the more confident LIGHTHOUSE is that there is
879 no such relation. Each of the three streams in module 2 (C) is trained to predict
880 scaled IC_{50} values with the use of BindingDB data. For instance, an interaction
881 score of 4 means that, if the compound has inhibitory activity, the IC_{50} would be
882 ~100 μM, whereas an interaction score of 9 means that, if the compound has
883 inhibitory activity, the IC_{50} would be ~1 nM. Note that module 2 only works if the
884 compound and protein interact, so this module is auxiliary to module 1. See also
885 Figures S1,S2, and Tables S1, S2, S3.
886

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

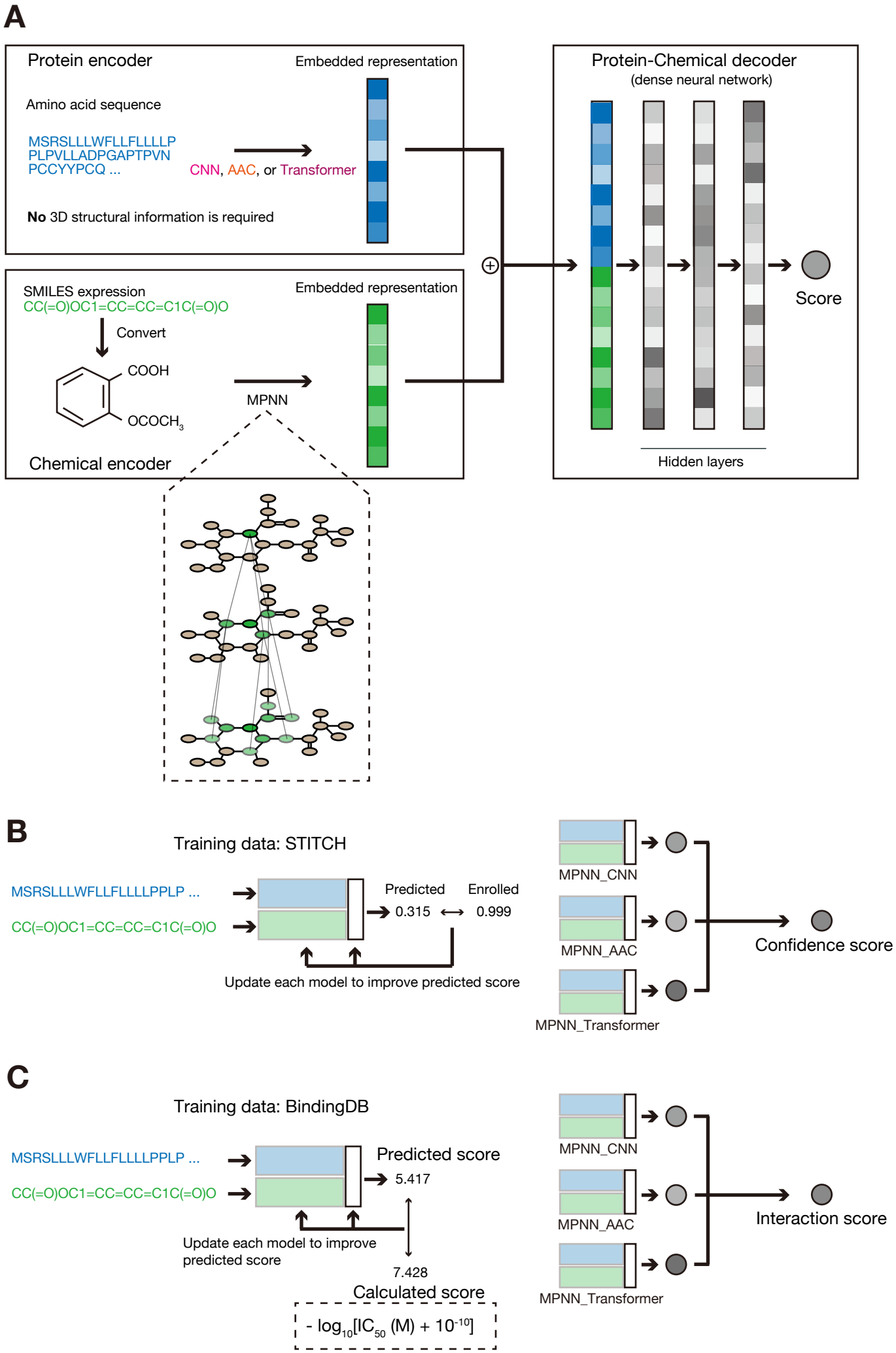
888 (A) For investigation of whether LIGHTHOUSE is able to enrich for compounds
889 with known targets, two data sets were generated from STITCH test data: a
890 "Positive" data set consisting of CPIs with the highest scores (>0.9), and a
891 "Negative" data set consisting of CPIs with the lowest scores (<0.2).
892 LIGHTHOUSE tended to yield higher confidence scores for CPIs in the "Positive"
893 data set.

894 (B) ROC curve showing that LIGHTHOUSE was able to distinguish the “Positive”
895 and “Negative” data sets in (A).
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 Figure S3.
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.

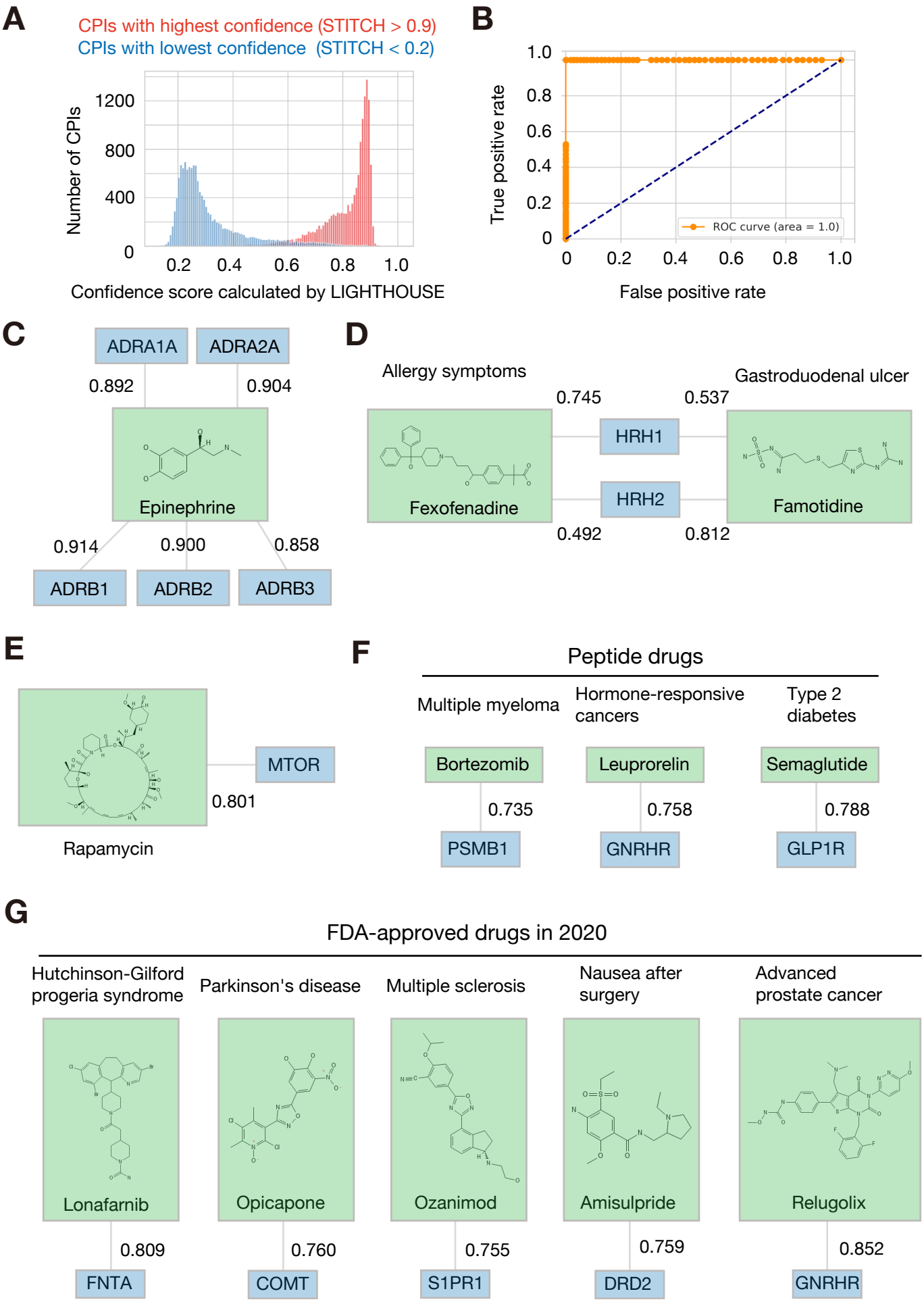
930 (E) *In vitro* PPAT activity assay performed in the presence of various
931 concentrations (1 nM, 5 nM, 10 nM, 50 nM, 100 nM, 500 nM, 1 μM, 5 μM, 10 μM,
932 50 μM, and 100 μM) of riboflavin 5′-monophosphate, with the determined IC_{50}
933 value being within the range predicted by LIGHTHOUSE. Data are shown for
934 four biological replicates. See also Figure S4.
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 Table S4.
981
982 **Figure 7. LIGHTHOUSE-based drug repurposing for COVID-19**
983 (A) Enrichment analysis of the top 500 potential statin targets identified in Figure
984 6 for COVID-19–associated gene sets. Minus $\log_{10}$-transformed q values are
985 shown.
986 (B) Prediction by LIGHTHOUSE of ethoxzolamide as a potential therapeutic for
987 SARS-CoV-2 infection on the basis of its confidence and interaction scores for
988 ACE2.
989 (C) Vero-TMPRSS2 cells were infected with wild-type SARS-CoV-2 at an MOI of
990 0.0001, cultured for 64 h in the presence of the indicated concentrations of
991 ethoxzolamide, and subjected to immunocytofluorescence analysis with
992 antibodies to SARS-CoV-2 N protein (green). Nuclei were stained with Hoechst
993 33342 (blue). Scale bars, 100 μm.
994 (D and E) Vero-TMPRSS2 cells challenged with wild-type (WK-521) or delta
995 strains of SARS-CoV-2, respectively, were cultured in the presence of various
996 concentrations of ethoxzolamide for 3 days and then subjected to the MTT assay
997 of cell viability. Nonchallenged cells were examined as a control. Data are
998 means $\pm$ SD for three independent experiments each performed in duplicate.
999 (F and G) Effect of ethoxzolamide on the SARS-CoV-2 load in culture
1000 supernatants of Vero-TMPRSS2 cells challenged with wild-type (WK-521) or
1001 delta strains of the virus, respectively. Data are from four independent

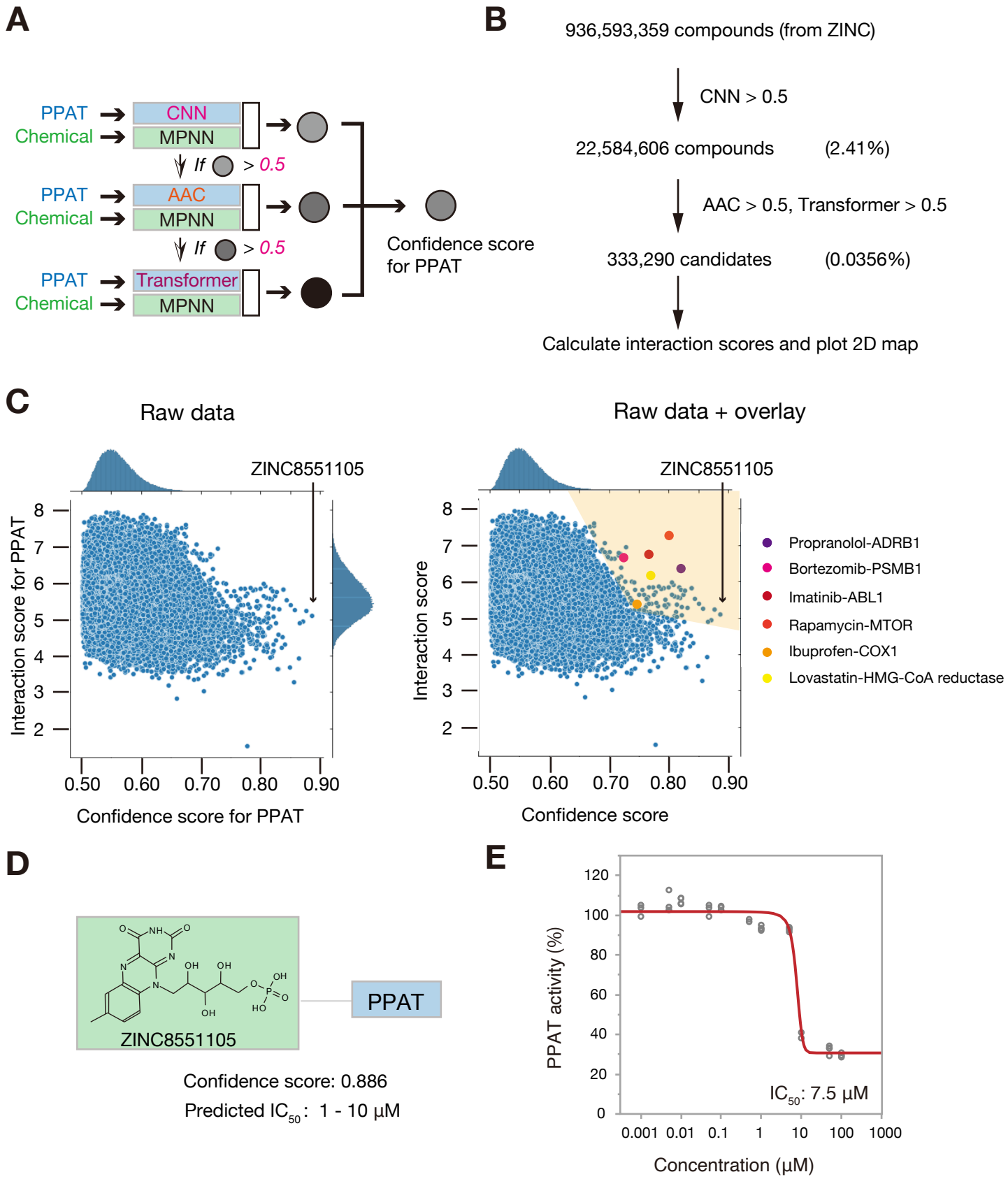
1002 experiments, with the graph line connecting mean values. $TCID_{50}$, median tissue
1003 culture infectious dose; N.D., not detected. See also Figures S5, S6, Table S5.



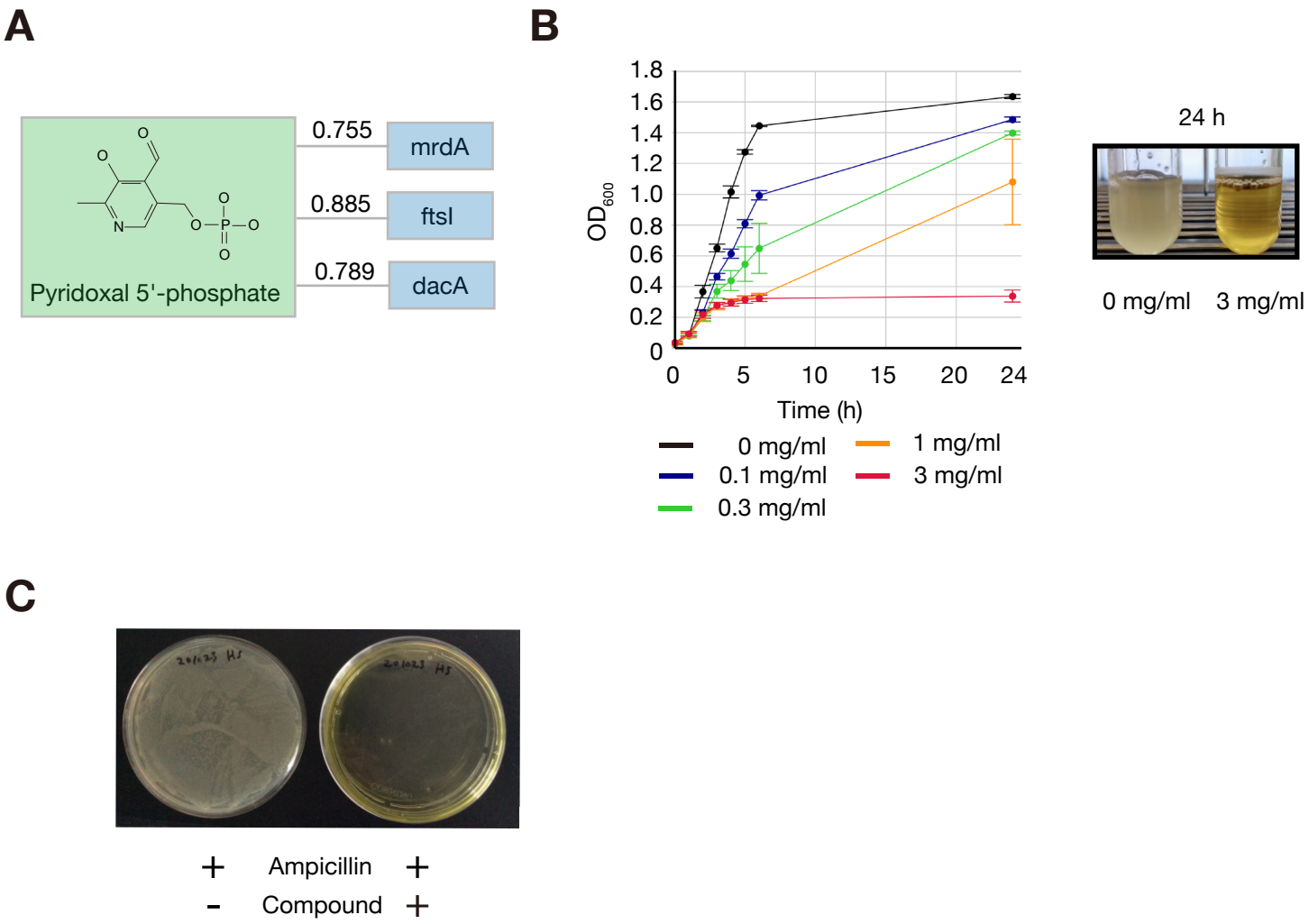
Shimizu et al., Figure 1



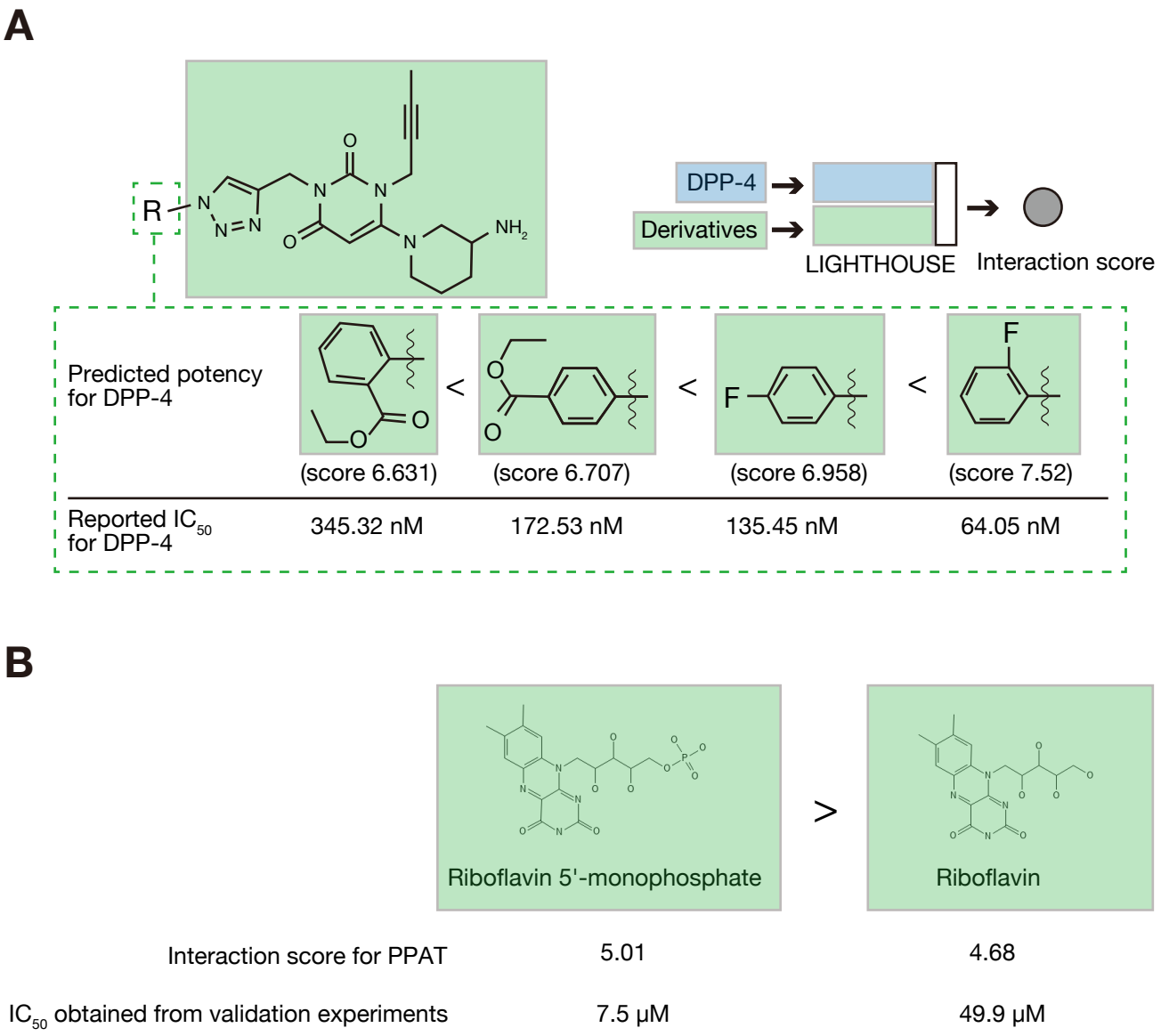
Shimizu et al., Figure 2



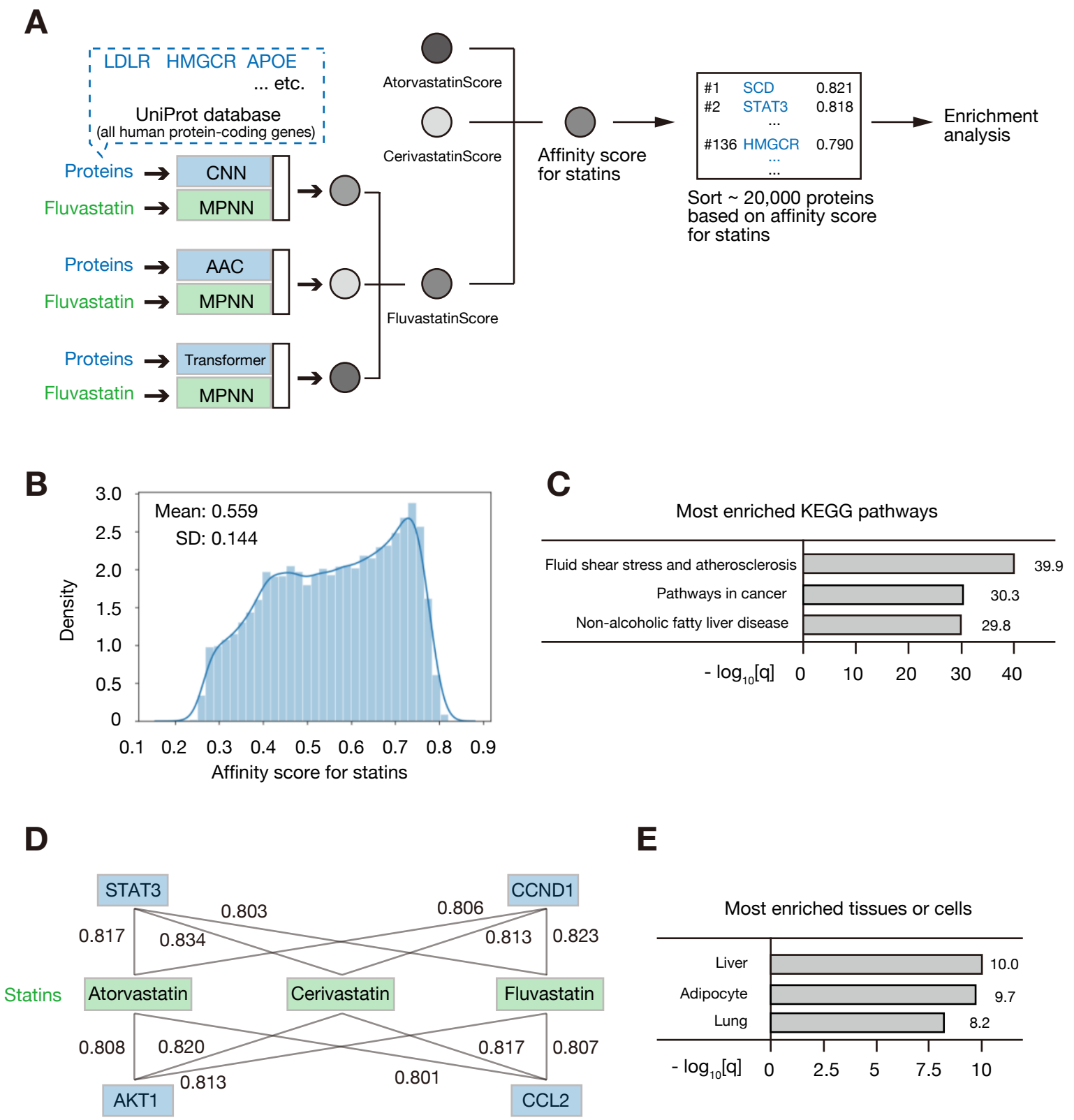
Shimizu et al., Figure 3



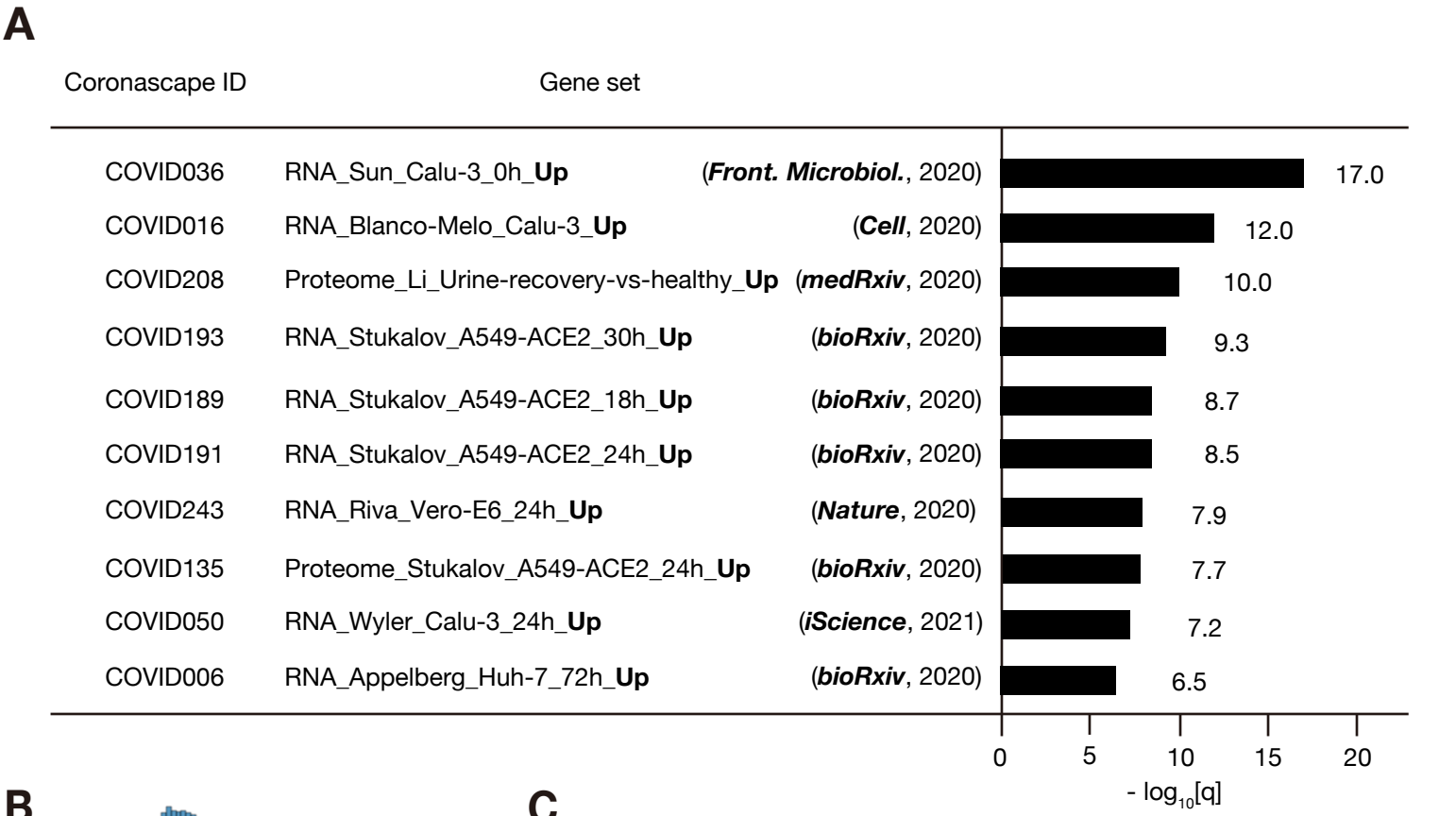
Shimizu et al., Figure 4

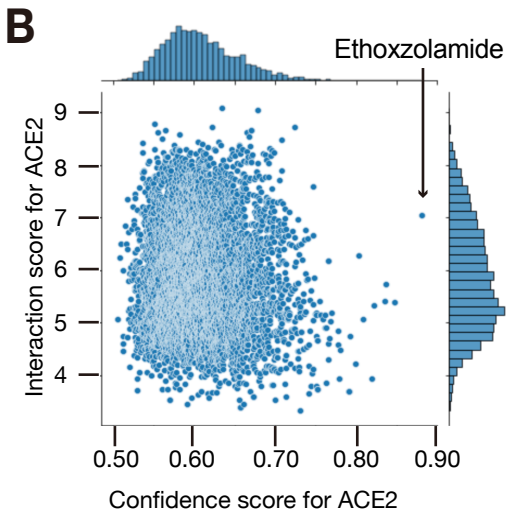


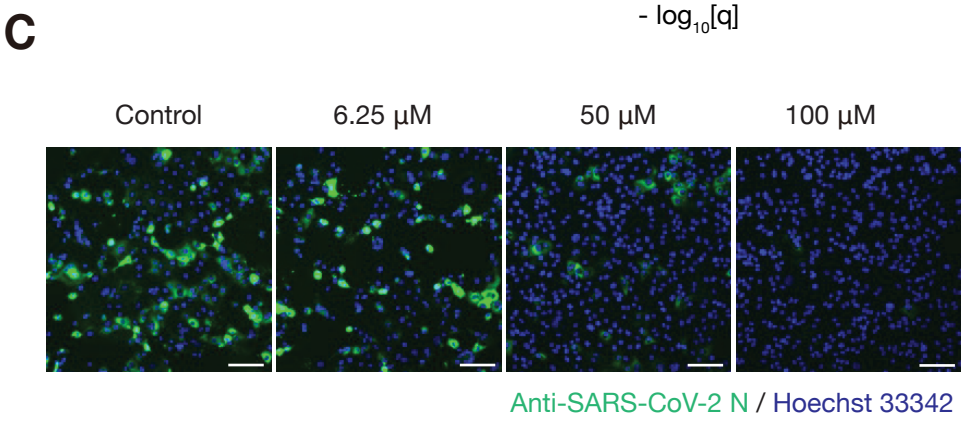
Shimizu et al., Figure 5

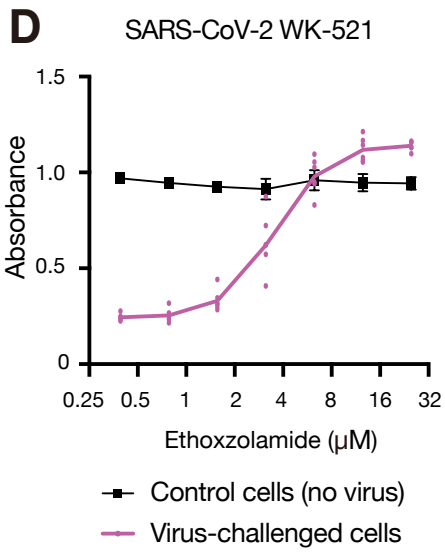


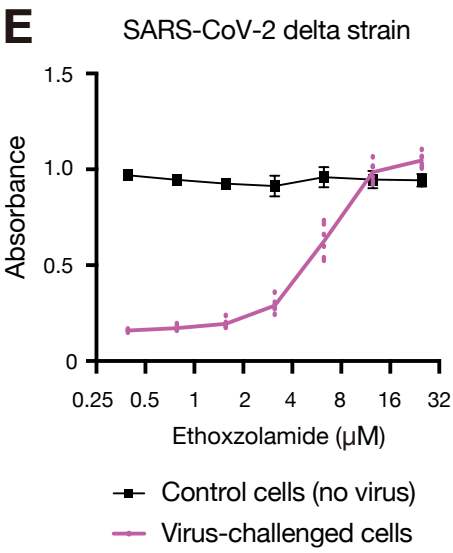
Shimizu et al., Figure 6

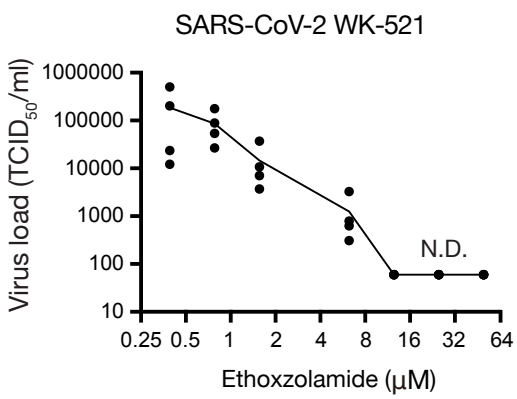


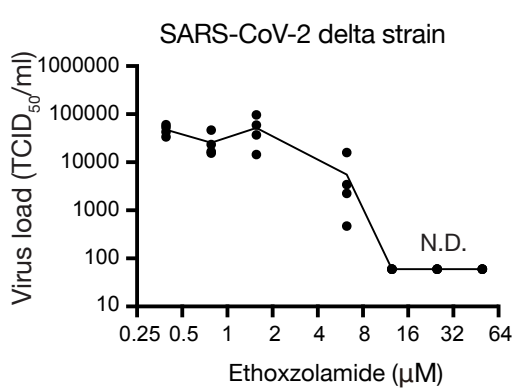












Shimizu et al., Figure 7