Discovery of novel quinoline papain-like protease inhibitors for COVID-19 through topology constrained molecular generative model

Jinsai Shang[1,2#*], Ting Ran[2#], Yongzhi Lu[1,2#], Qi Yang[1,3#], Guihua Zhang[2#], Peiqi Zhou[2], Wenqi Li[2], Minyuan Xu[2], Minxian Dai[2], Jinpeng Zhong[2], Hua Chen[2], Pan He[2], Anqi Zhou[1], Bao Xue[2], Jiayi Chen[2], Jiyun Zhang[2], Kunzhong Wu[2], Xinyu Wu[2], Miru Tang[2], Xinwen Chen[2,3*] and Hongming Chen[1,2,4*]

Affiliations

[1]School of Basic Medical Sciences, Guangzhou Laboratory, Guangzhou Medical University, Guangzhou 511436, China.

[2]Department of Drug and Vaccine Discovery, Guangzhou National Laboratory, Guangzhou 510005, China.

[3]State Key Laboratory of Respiratory Disease, Guangzhou Medical University, Guangzhou, 511436, China.

[4]Lead contact

# These authors contributed equally to this work

* Correspondence: Jinsai Shang: shang_jinsai@gzlab.ac.cn; Xinwen Chen: chen_xinwen@gzlab.ac.cn; Hongming Chen: chen_hongming@gzlab.ac.cn

Abstract

Papain-like protease (PL^{pro}) plays a critical role in both viral polyprotein processing and host antiviral immune suppression in SARS-CoV-2 infection, which causes COVID-19. Although several drugs have been approved for COVID-19, such as Remdesivir, Nirmatrelvir etc., none of the PL^{pro} inhibitors have been approved for the treatment of COVID-19. The advent of artificial intelligence-based drug design methods has significantly accelerated the process of drug discovery. In current study, by harnessing the power of a topology constrained molecular generative model, we discovered a novel series of PL^{pro} inhibitors with strong potency against prevalent SARS-CoV-2 variants. Following a structure based computational approach for optimization, our lead compound, GZNL-2002, achieved decent PL^{pro} inhibitory potency and

30 favorable pharmacokinetic properties, which warrants further development as a potential candidate
31 compound for COVID-19 disease.
32

---

33 **Introduction**

34 Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is the infectious virus
35 responsible for corona virus disease 2019 (COVID-19). It was initially discovered in China in
36 December 2019 and has rapidly evolved into a global epidemic with a significant cost to human
37 society. Over the past four years, scientists worldwide have been exploring various approaches to
38 treat infected individuals and reduce the severity of the disease. To date, several COVID-19
39 vaccines have been developed and granted emergency authorization for market distribution(*1*). All
40 these vaccines target the spike protein to prevent the virus from entering the host cells. However,
41 the high mutation rate of the spike protein, especially within the receptor-binding domain (RBD),
42 diminishes the long-term protective effectiveness of the vaccines(*2*). This underscores the need to
43 develop alternative therapeutic modalities, particularly small molecule antiviral drugs that target
44 more conserved viral proteins.

45 Currently, the FDA has approved three small-molecule drugs, including remdesivir,
46 molnupiravir and nirmatrelvir(*3*). They all target relatively conserved proteins across different
47 SARS-CoV-2 variants. The first two drugs are viral RNA-dependent RNA polymerase (RdRp)
48 inhibitors, while the third one is an inhibitor of the viral main protease (M^{pro}), also known as 3CL
49 protease ($3CL^{pro}$). Remdesivir was the first drug authorized for emergency use, but its
50 administration is limited to intravenous infusion for hospitalized patients with severe symptoms(*4,*
51 *5*). Its clinical efficacy has been found to be sub-optimal, as demonstrated by several clinic trials(*6*).
52 Similar to remdesivir, molnupiravir is a nucleotide analogue and functions as a mutagen, inhibiting
53 SARS-CoV-2 replication by increasing the viral mutation rate(*7*). However, there are concerns
54 that molnupiravir may also induce similar mutations in the host(*8*). Nirmatrelvir, combined with a
55 cytochrome P450 3A (CYP3A) inhibitor (*i.e.,* ritonavir), is an oral medication and suitable for
56 non-hospitalized patients(*9*). Ensiltrelvir, a M^{pro} inhibitor approved in Japan(*10*), has demonstrated
57 high efficacy in clinical trials. However, it is also a potent CYP3A inhibitor, which may lead to
58 severe adverse drug-drug interactions with other medications(*11*). Recently, mutations-induced
59 drug resistance have been identified in cell cultures and COVID-19 patients treated with these
60 drugs(*12, 13*). Therefore, new antivirals with alternative mechanisms are still urgently needed.

61 The papain-like protease (PL^{pro}) is one of the two viral proteases responsible for cleaving the
62 polyproteins encoded by the SARS-COV-2 genome(*14*). Additionally, it can remove the ISG-15
63 and ubiquitin modifications on host proteins(*15*). This dual enzymatic function allows it to both
64 promote virus replication and suppress the host's immune response. Moreover, PL^{pro} is a highly
65 conserved protein across SARS-CoV-2 variants(*16*), making it a promising drug target for anti-
66 COVID-19 therapy. The catalytic site of the protein is situated within a small buried pocket,
67 adjacent to an induced substrate-binding cavity by the reshaping of the BL2 loop, a flexible beta
68 turn/loop region (Gly266–Gly271)(*17*).
69 **GRL0617** was originally discovered as a non-covalent anti-SARS-CoV PL^{pro} inhibitor
70 binding at the induced substrate-binding cavity and later validated to have inhibitory potency
71 against SARS-COV-2 PL^{pro} at micromolar level(*17*). Analogous of **GRL0617** were developed
72 with moderate enzymatic and cellular activities, such as **Jun9-84-3**(*18*), **XR8-23**(*19*),
73 **Jun11273**(*20*) and covalent inhibitors **CP7**(*21*). Recently, **Jun12862** was reported to have in vivo
74 efficacy in SARS-CoV-2 infected mice, along with oral bio-availability(*22*). **GZNL-P4** is a potent
75 compound identified in our previous study (*23*), but structural optimization is needed to improve
76 its low liver stability. The development of PL^{pro} inhibitors at the clinical stage remains rare. To
77 date, only one PL^{pro} inhibitor has been reported to have entered clinical trials in China, and its
78 structure has yet to be disclosed(*24*). Therefore, the development of PL^{pro} inhibitors remains
79 necessary to explore their potential as COVID-19 therapy.
80 Artificial intelligence (AI) technology has shown great potential to accelerate drug
81 discovery(*25*). Among these models, deep molecular generation models have garnered particular
82 attention due to their ability to autonomously produce novel molecular structures. Early molecular
83 generation models primarily involved random sampling from the learned continuous molecular
84 chemical space to generate a large number of molecular structures(*26*). However, due to the
85 vastness of chemical space, the probability of discovering active molecules through sampling these
86 models is relatively low. In recent years, there has been increasing focus on achieving goal-directed
87 molecular design through the development of conditional molecular generation models. Key
88 approaches include altering the form of model input(*27*), utilizing conditional deep neural
89 networks (*28*), applying transfer learning(*29*), and employing reinforcement learning(*30*) etc.
90 Some of the latest methods directly decode the learned three-dimensional structural features of

91 protein pockets into molecular chemical structures (*31*). Although numerous methods have been
92 developed, only a few of them have been reported with application in real-world scenario.
93 In earlier research, we developed Tree-Invent(*32*), a graph-based molecular generation model
94 with a topology constraints. This model is versatile across different drug design scenarios and
95 excels in achieving precise control over structural topology during molecular generation. In this
96 model, a molecular graph is represented as a topological tree comprising ring nodes, non-ring
97 nodes, and edges connecting the nodes. Molecule generation is achieved by carrying out operations
98 of node addition, ring generation, and node connection in an autoregressive manner.
99 Reinforcement learning (RL)(*33*) is also integrated within the Tree-Invent model to facilitate the
100 exploration of chemical space toward desired properties. We applied this method to conduct
101 scaffold hopping exercise and discovered a novel quinolone chemical series exhibit nanomolar
102 potency in enzymatic assay and excellent anti-viral potency in cellular experiments. Particularly,
103 our best compound demonstrates reasonable in vitro and in vivo oral PK properties. The significant
104 selectivity over normal human cells and weak inhibition on the CYP and hERG proteins indicate
105 its promising safety profile for antiviral treatments.
106
107 **Results**
108 **Scaffold hopping through topology constrained molecular generation with Tree-Invent**
109 So far, exploration of the structure-activity relationship (SAR) based on the structure of **GRL-**
110 **0617** has resulted in several potent PL^{pro} inhibitors (Fig.1A). In current study, we utilized the Tree-
111 Invent model to design novel PL^{pro} inhibitors, using the crystal structure of PL^{pro} enzymatic domain
112 in complex with known inhibitor GRL0617, which exhibits potency at micromolar range. Its
113 crystal structure (PDB ID: 7JRN) is shown in Fig.1B, where the naphthalene ring serves as the
114 head group, is located at the BL2 groove sandwiched between the aromatic residue Y269 and the
115 hydrophobic residues P247 and P248. The aniline group in the tail group is located at the P3 sub-
116 pocket and extends into the P2 sub-pocket near L162 through a methyl group. The linker group is
117 located at the P4 sub-pocket, where the hydrogen bonds with D164 and Q269 are formed by the
118 amide group. Accordingly, the structure of **GRL0617** can be divided into three parts, namely the
119 head, linker and tail part, corresponding to three sub-pockets in the substrate binding site (Fig.1C)
120 (*17*).

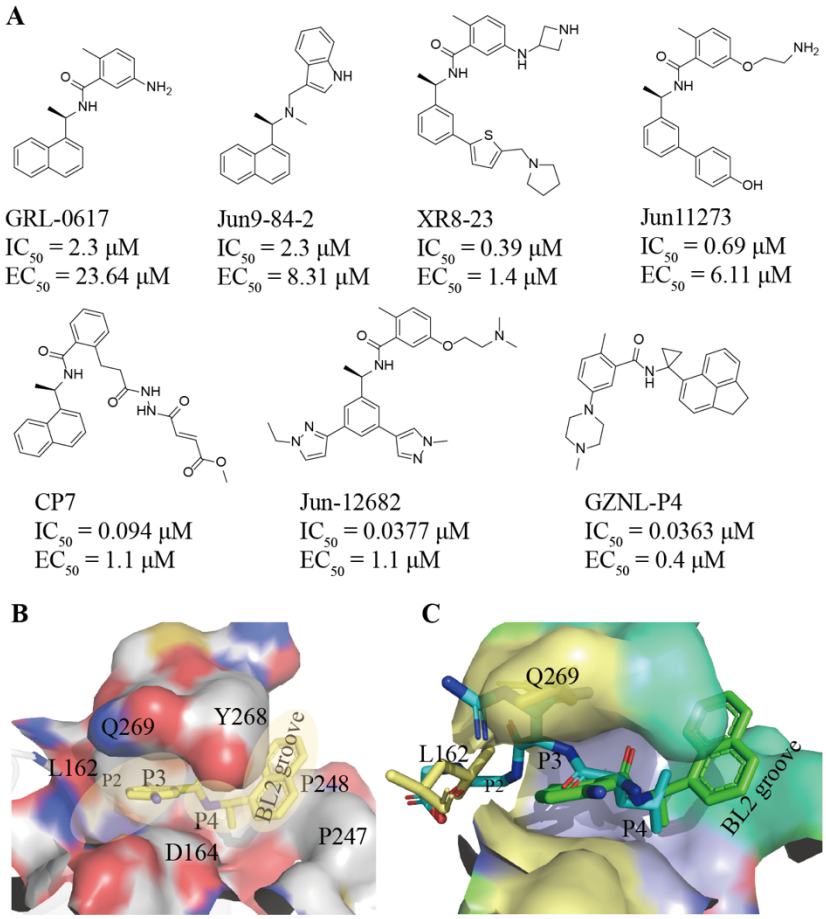


Fig.1. Structure and binding mode of available PL^{pro} inhibitors. A) **Structure of available SARS-Cov-2 PL^{pro} inhibitors**. IC_{50} and EC_{50} are enzymatic and antiviral activity for wild-type PL^{pro}, respectively. B) Residues of PL^{pro} involved in the interaction with **GRL0617 (**PDB ID: 7JRN**)**. The binding pocket was shown as protein surface and colored by element types. Ligand was shown as yellow sticks. Dashed circles represent the sub-pockets. C) Overlap between **GRL0617** and substrate peptide (LRGG) in the active pocket of PL^{pro}. The substrate peptide is extracted from the co-crystal structure of SARS-COV PL^{pro} with ubiquitin (PDB ID: 4M0W)(*34*). The BL2 groove, P4 and P3 sub-pockets are shown as green, blue and yellow surface respectively. Green and cyan sticks refer to the **GRL0617** and substrate peptide (sequence LRGG). The residues L162 and Q269 are shown as yellow sticks.

In our previous work (*23*), we identified a potent compound, **GZNL-P4**, which features a tricyclic dihydroacenaphthylene (DHAN) scaffold. The co-crystal structure of **GZNL-P4** with PL^{pro} (PDB ID: 8YX2, to be published) reveals that its tail group extends into the solvent-exposed region at the P3 sub-pocket. The positively charged piperazine nitrogen in this group can form salt bridge with E167, an interaction absent in the **GRL0617** complex structure (Fig.2A). However, **GZNL-P4** has poor liver metabolic stability (rat liver microsome stability $t_{1/2}$ 7.95 min) (*23*), which is believed to be associated with the head DHAN group as observed in the metabolism of **GRL0617** (*35*), so finding other head groups is needed to address the stability issue.

140 Herein, we employed Tree-Invent model to explore chemical space of head group. The
141 structure of **GZNL-P4** was transformed to a topology tree with four nodes connected by single
142 bonds (Fig.2B). To focus on the generation of the head group, node N1 was predefined as the
143 substructure comprising the linker and tail part of **GZNL-P4** and kept unchanged in the process
144 of molecule generation. Node N2 was designated as a changeable node grown from node A to
145 replace the head group of **GZNL-P4**. Analysis of the binding modes of **GZNL-P4** and **GRL0617**
146 indicates that the aromatic ring system of the head group is important for binding through π-π
147 stacking interaction with Y268 and hydrophobic interaction with P247 and P248 (*17*). To mimic
148 this binding mode, node N2 was confined to contain 5,6- or 6,6-fused bicyclic systems with at
149 least one aromatic ring. Furthermore, two nodes, *i.e.,* N3 and N4, were additionally appended to
150 node N2 as R-group decorations. However, their generation was non-essential and entirely
151 contingent upon the model's decision regarding node addition. With these topological constraints,
152 Tree-Invent was implemented with RL to generate molecules (Fig.2C), and two rounds of RL were
153 conducted respectively for the generation of 5,6- and 6,6-fused bicyclic molecules. The crystal
154 structure of **GRL0617** was chosen as the protein model for carrying out docking during the
155 generative process and the docking score was transformed into reward score during the RL process,
156 aiming to generate molecules with high affinity.
157 Through 1000 epochs of RL, the learning curves of the Tree-Invent model converge (Fig.S1A),
158 suggesting the model has learned how to generate high-score molecules with 5,6- and 6,6-fused
159 bicyclic systems. Combining the generation sets of the two RL runs, 2817 unique molecules were
160 obtained. Among them, 1270 molecules have 5,6-fused ring structure and 1547 molecules have
161 6,6-fused ring structure (Fig.S1B). Moreover, the substructures corresponding to node N1 in these
162 molecules are kept the same as in **GZNL-P4**. Half of the molecules have naked ring structure in
163 the head group, while the others have no more than two substitutes on the ring structure as expected
164 (Fig.S1C). This demonstrates that the Tree-Invent model is capable of generating molecules
165 satisfying the topological constraints. It was also found that the naphthalene group of **GRL0617**
166 can be reproduced which proves the capability of this method in identifying active molecules.
167 Subsequently, all generated molecules were redocked to the binding pocket using the Glide
168 docking program under the standard precision mode (Glide-SP)(*36*), resulting to 367 molecules
169 with similar binding mode to **GZNL-P4** and **GRL0617** and docking scores lower than -7.0
170 kcal/mol, a cutoff to distinguish suitable molecules from all generated molecules (Fig.S1D).

Moreover, these remaining molecules are all compliant with the Lipinski rules(*37*). Chemical space analysis reveals that these molecules can be grouped into several categories, and each cluster has some molecules with good docking scores (Fig.2D).

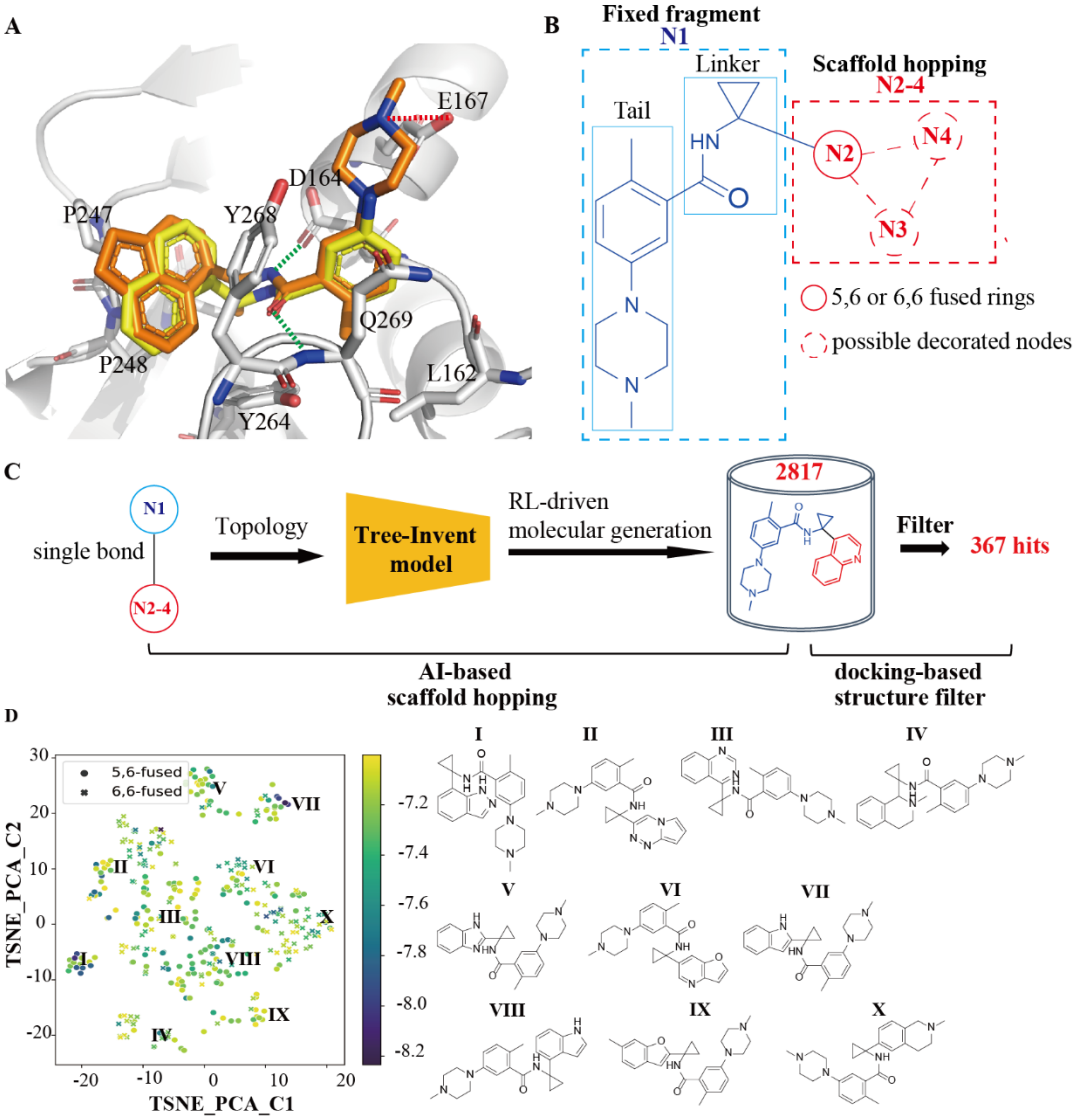


Fig.2. **Scaffold hopping of the structure of GZNL-P4.** A) Binding mode of **GZNL-P4**. The crystal structures of **GZNL-P4** (orange) and **GRL-0617** (yellow) are overlapped. Ligands and residues surrounding the binding pockets are shown as sticks. Protein secondary structure is shown as transparent cartoons. Green and red dashes represent hydrogen bond and salt bridge interactions. B) Transformation of **GZNL-P4** to a topological tree. The solid and dashed circles represent essential and inessential nodes for molecular generation respectively. The dashed lines represent the bonds to the inessential nodes. C) Workflow of scaffold hopping of the **GZNL-P4** structure based on the Tree-Invent model using the RL strategy. The blue and red circles represent prefixed and changeable nodes. D) Chemical space analysis based on the PCA analysis (*38*) using the Morgan fingerprint (*39*). tSNE plot (*40*) was used to display the chemical space of molecules filtered based on docking scores (represented by the color of the heatmap) and their pharmacophore model. The labels I-X indicate ten diverse structures with reasonable docking scores.

187 **Identification of novel scaffold of PL^{pro} inhibitors**

188 Considering synthesis feasibility, nine molecules were chosen for synthesis. The synthetic
189 route was illustrated as Scheme 1 in the supplemental text. The structures of the synthesized
190 compounds are shown in Fig.3A. Their docking poses can be found in Fig.S2. Then, they were
191 tested using a fluorescence-based biochemical assay(*41*) to evaluate their inhibitory activities on
192 PL^{pro} (Fig.3B). Compounds **1** and **2**, containing a quinoline ring in the head group, are the most
193 active compounds. Their aromatic nitrogen atoms in the quinoline ring are situated at the para-
194 position to the linker substituent. Chlorination on the quinoline ring in compound **1** shows a minor
195 impact on the activity, while compound **2** is suggested to have better binding affinity (Fig.3C).
196 Compound **6** with benzothiophene exhibits somewhat lower activity than compound **2**. Other
197 compounds having 5,6 ring systems such as indole, benzopyrazole and 1H-pyrrolopyridine have
198 much worse activity, regardless of whether the head group is a five-membered or six-membered
199 ring. Among them, compound **8** with azaindole is the weakest. N-methyl substitution on the indole
200 ring of compound **9** improves the activity, probably because it makes additional van der Waal
201 interaction with the residue P248. Compound **5** with furopyridine has a special substitution
202 position on the head group, indicating that the direction of substitute on the head group is very
203 important for activity. Through the AI-driven R-group generation process, we successfully
204 identified several PL^{pro} inhibitors with new scaffolds having reasonable potency. The liver
205 metabolic stability of compound **2** was measured, and to our delight, its half-life in liver microsome
206 has been largely improved compared to **GZNL-P4** as shown in Fig.6E.

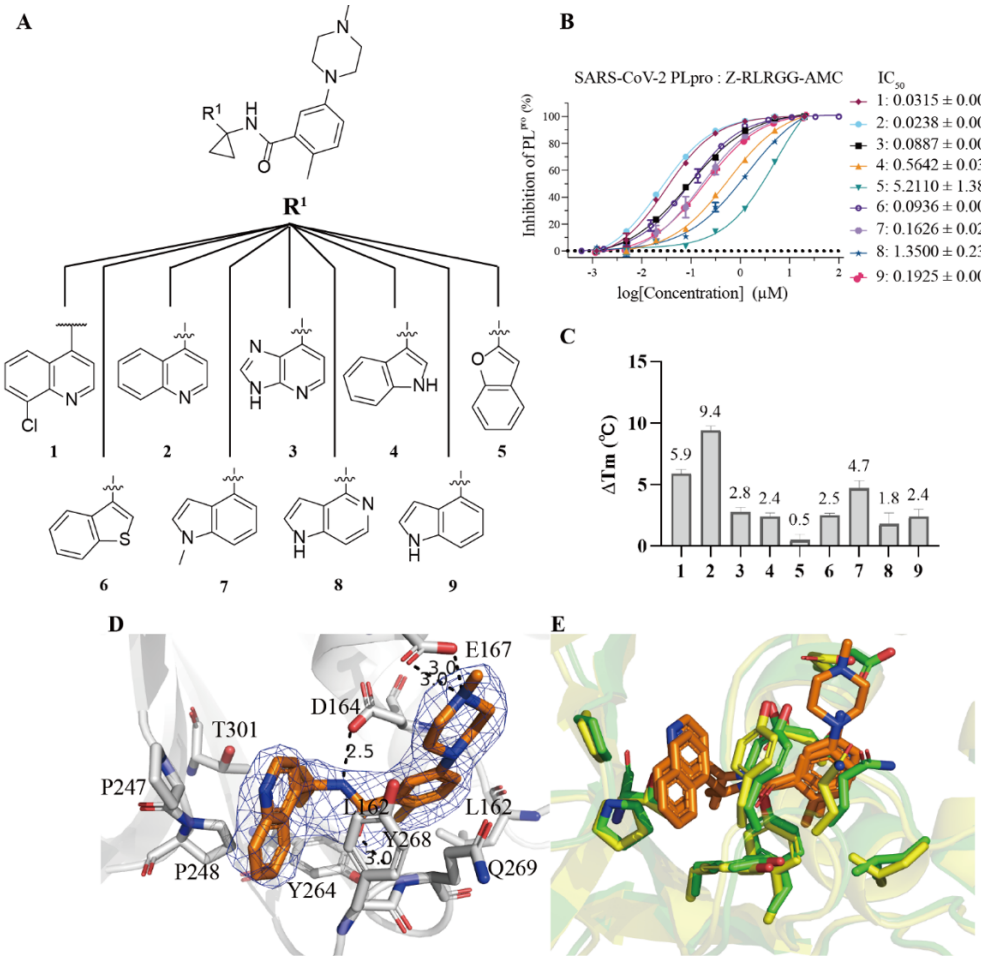


Fig.3. Lead compound identification. A) Structures of nine synthesized compounds **1-9**. R^1 groups are shown at the top. B) Dose-dependent inhibitory activity of compounds **1-9** against PL^{pro} as determined by the fluorescence-based biochemical assay. Half maximum inhibitory concentration (IC_{50}) is shown in the legend. Data are presented as mean ± s.d. of three technical replicates. C) Differential scanning fluorimetry assay analysis of SARS-CoV-2 PL^{pro} with compounds. The error bars are mean ± s.d. of three replicates. D) The binding mode of compound **2** was resolved by X-ray diffraction at 2.8 Å (PDB code: 8ZSE). Compound **2** is shown as orange sticks. Residues around the binding pocket are shown as gray sticks. Protein secondary structure is shown as gray cartoons. Black dashes represent hydrogen bond and salt bridge. E) Alignment of the crystal structures with compound **2** (yellow) and **GRL0617** (green, PDB code: 7JRN). Protein secondary structure is shown as cartoons and residues around the binding pocket are shown as sticks. Ligands are specifically shown as orange sticks.

To better understand the binding mode of compound **2**, we resolved its co-crystal structure in complex with PL^{pro} (Fig.3D). The residues surrounding the binding pocket maintain nearly identical conformations to those observed in the co-crystal structure of **GRL0617**, and there is a good alignment between the binding poses of the two compounds (Fig.3E). The binding mode analysis shows that the quinoline ring of compound **2** is overlapped with the head group of **GRL0617**. As expected, the quinoline ring makes π-π stacking interaction with Y268, and extend the interaction with P247 and P248. The nitrogen atom on the quinoline ring is exposed to solvent,

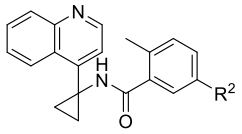
226 making no obvious binding with protein residues. The cyclopropyl group binds at the bottom of
227 the pocket surrounding Y264 and T301, and the amide linker group makes hydrogen bonds with
228 D164 and Q269. The tail group, i.e., piperazine-substituted benzyl group, is exposed to solvent,
229 and the positively charged piperazine nitrogen in this group can form salt bridge with E167, an
230 interaction absent in the GRL0617 complex structure, while the benzyl group extends into the sub-
231 pocket near L162 through a methyl group. This binding mode is consistent with that of **GZNL-P4**,
232 which provides clear guidance for structural optimization.
233
234 **Structural optimization of compound 2**
235 To further improve the potency of compound **2**, we explored substituents at the R^2 position
236 to optimize its interaction with E167. We selected commercially available amine-containing
237 reagents to replace the N-methylpiperazine group of compound **2** and designed a virtual library of
238 2347 compounds containing positively charged amine groups (Fig.S3). Molecular docking using
239 the co-crystal structure of compound **2** as docking model was then carried out, compounds whose
240 docking pose is similar to the binding pose of compound **2** and simultaneously maintains the salt
241 bridge with E167 in an appropriate distance (the cutoff is set to 4.0 Å) were selected. At last, 29
242 top-ranked compounds (by Glide docking score) were selected to synthesize (Scheme 2 in
243 supplemental text) and their structures are listed in Table 1. They can be divided into two types in
244 terms of the ring count in the R^2 group.
245 Following SAR analysis reveals that the R^2 group greatly influences the inhibitory activity of
246 PL^{pro} (Table 1 and Fig.S4). Several compounds show superior activity to compound 2, and three
247 compounds demonstrate IC_{50} values lower than 0.015 μM. Among type I molecules with a single
248 ring in the R^2 group, compound **10** with piperazine substitute has worse potency than compound
249 **2** having a methylpiperazine group, which means that the activity is sensitive to the position of
250 methyl substituent on the piperazine ring. Compound **2** with 4-methylpiperazine exhibits much
251 higher activity than compounds **14** and **15** which contain the substitutes of 2- and 3-
252 methylpiperazine. The compounds with hydroxyalkyl groups substituted on the 4-position of
253 piperazine have slightly worse activity than compound **2**. However, substituting with
254 dimethylaminoalkyl groups, as seen in compound **12**, negatively impacts activity. N-sulfonylation
255 of piperazine significantly reduces activity, likely due to the disruption of the salt bridge formed
256 between piperazine and the acid group in E167. It is worthwhile noting that the bridged piperazine

257 significantly enhances activity. Compound **16** is the most potent in this series with IC_{50} of 7.1 nM.
258 Compounds bearing similar cycloamine groups, including the ones with spiral and fused rings,
259 also demonstrate reasonable activity, although don't get improvement compared to compound **2**.
260 These groups can mimic the salt bridge interaction of piperazine. Nevertheless, the introduction of
261 seven-member-ring results in a reduction of activity, which may be due to the sub-optimal salt
262 bridge interaction of the secondary amine in the enlarged ring.

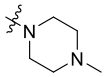
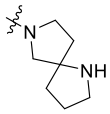
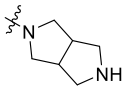
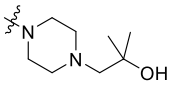
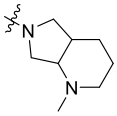
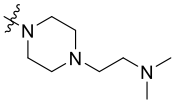
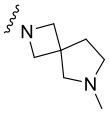
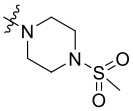
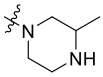
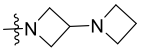
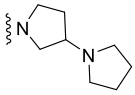
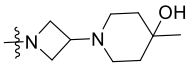
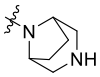
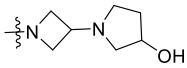
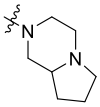
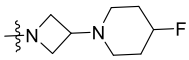
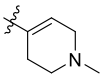
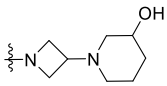
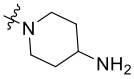
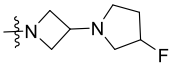
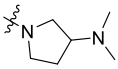
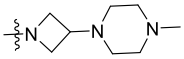
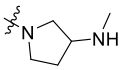
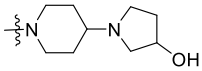
263 Type II molecules contain the R^2 substituents involving two cycloamine groups linked by a
264 single bond. Interestingly, the majority of compounds exhibited comparable activity to compound
265 **2**, except for compounds **35** and **36**. The reduced activity of compound **35** is attributed to the N-
266 methyl-piperazine substituent, which may form a sub-optimal salt bridge with E167. The weak
267 activity observed for compound **36** indicates that piperidine is worse than piperazine for the
268 substitution on the benzene ring, as evidenced by the decrease in the activity of compound **20**. On
269 the other hand, compounds featuring azetidines substituted with pentacyclic/hexacyclic amines
270 demonstrated the highest activity within the sub-group. Notably, compounds **30** and **31** exhibit
271 potency comparable to compound 16, making them also promising candidates for further profiling.

272 As shown in Fig.4A-C, the docking poses of the most potent compounds, compounds **16**, **30**,
273 and **31,** are consistent with that of compound **2**, which is characterized by the sandwiched
274 conformation of the quinoline ring, the hydrogen bonds with D164 and Q269, and the salt-bridge
275 with E167. Furthermore, compounds **30** and **31** both made an additional hydrogen bond with E167,
276 which may explain the higher potency of these compounds than compound **2**. In the binding pose
277 of compound **16**, the bridged piperazine also makes a hydrogen bond with E167 in addition to
278 forming a salt bridge. Moreover, the bridged carbon seems to lock the piperazine in a favorable
279 conformation for interacting with E167, which may explain the superior performance of compound
280 **16** to other compounds.

Table 1. SAR exploration of R^2 group



Compound	R^2	PL^{pro} IC_{50} (nM)	Compound	R^2	PL^{pro} IC_{50} (nM)

2		0.0238±0.0011	**23**		0.0992±0.0217
10		0.0327±0.0016	**24**		0.0563±0.0059
11		0.0288±0.0018	**25**		0.0156±0.0018
12		2.2330±0.2485	**26**		0.2861±0.0224
13		1.0300±0.1295	**27**		0.3016±0.0287
14		0.0453±0.0069	**28**		0.0376±0.0055
15		11.0700± 14.6380	**29**		0.0256±0.0050
16		**0.0071±0.0022**	**30**		**0.0108±0.0036**
17		0.0316±0.0065	**31**		**0.0131±0.0015**
18		0.0520±0.0088	**32**		0.0468±0.0113
19		0.0284±0.0022	**33**		0.0494±0.0019
20		0.3836±0.0510	**34**		0.0519±0.0107
21		0.0483±0.0107	**35**		0.1817±0.0131
22		0.0560±0.0090	**36**		1.0540±0.1536

281

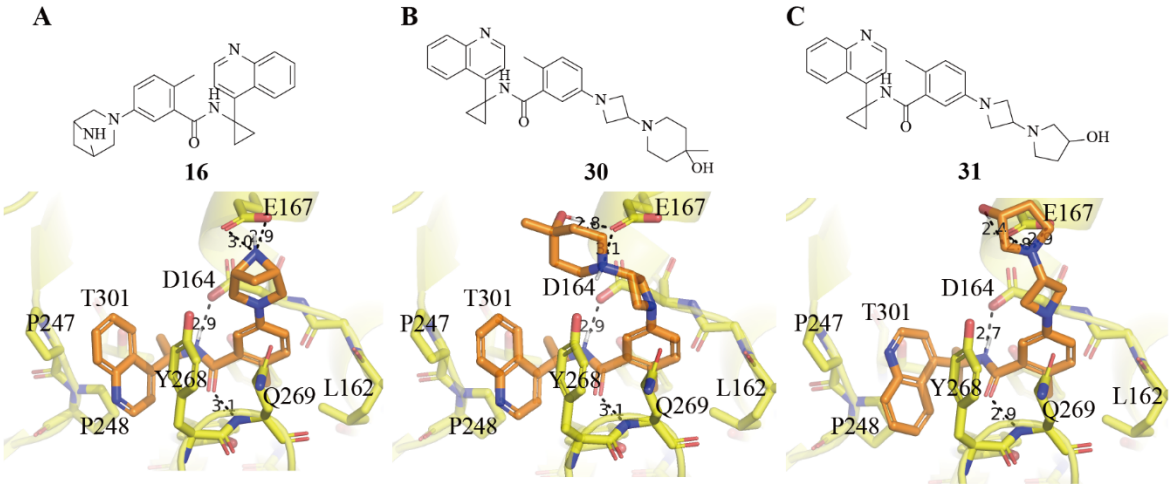


Fig.4. Binding mode of active compounds. Docking poses of compound **16** (A), **30** (B), **31** (C). Ligands and residues are shown as orange and yellow sticks respectively. Black dash represents either hydrogen bond or salt bridge interactions labeled with the distance.

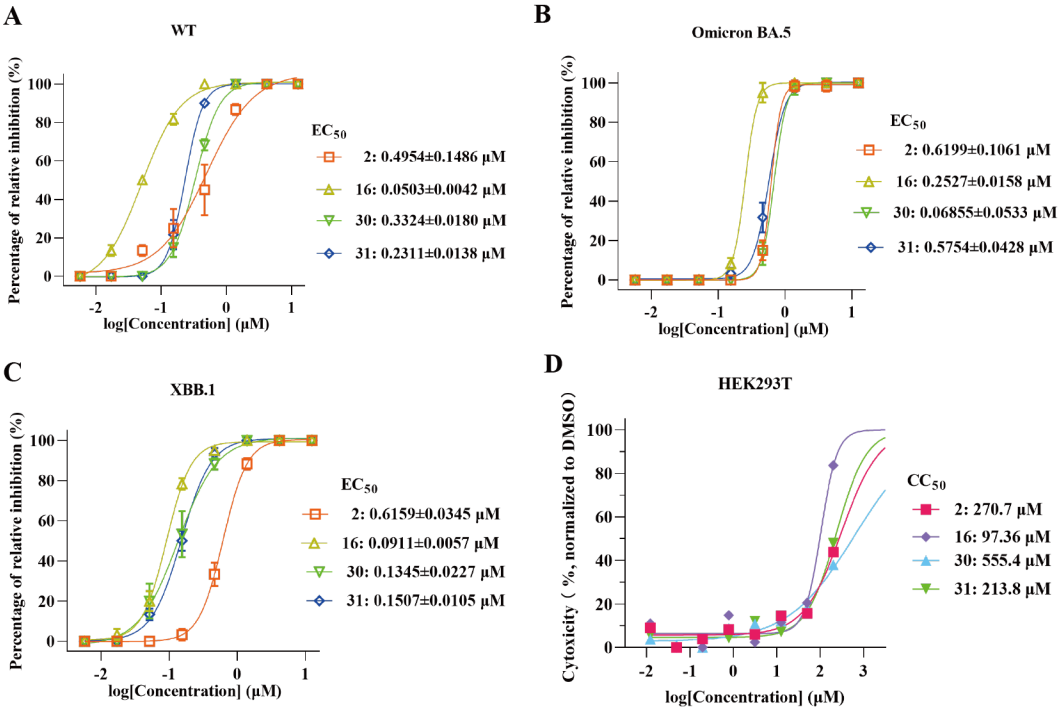
Antiviral activity, liver stability and PK profiling

The antiviral activities of several potent compounds were measured using cell-based assay. As shown in Fig.5A-C, the antiviral potency with 50% effective concentration (EC_{50}) of compound 2 against wild-type SARS-CoV-2 and two epidemic variants Omicron BA.5 and XBB.1 in infected Vero E6 cells are 0.4954 ± 0.1486 μM, 0.6199 ± 0.1061 μM and 0.6159 ± 0.0345 μM. At the same time, compound **2** shows rather low toxicity to normal cell line of human (CC_{50} = 270.7 μM) (Fig.5D). More importantly, the liver stability of Quinoline series in both rat and human liver microsomes was significantly improved comparing with **GZNL-P4** (Fig.5E).

Structural optimization on compound **2** resulted in three more potent compounds, compound **16**, **30** and **31**, which have enhanced anti-viral activity and large enough safety window (CC_{50}/EC_{50} > 370 folds for all variants). Notably, compound **16** has the best antiviral activity compared to compound **30** and **31** on all three tested virus variants. And, the variant of Omic-BA.5 is least sensitive to all three PL^{pro} inhibitors.

In general, compound **16** has the best inhibition of enzyme and virus as well as excellent liver metabolic stability, it was selected as the lead compound for further profiling (Fig.5F). The CYP and hERG inhibition at 10 μM and 30 μM for compound **16** were summarized in Table 6, the results indicate the compound has low risk in inhibiting selected CYP isoforms and hERG channel. The oral in vivo PK properties of compound **16** at a dose of 20 mg/kg is also listed in Table 6 and

305 the results show that compound **16** (here renamed as **GZNL-2002**) is a promising lead compound
306 suitable for further optimization.
307



E

parameter	2	16	30	31
Rat Livermicrosome stability $t_{1/2}$ (min)	45.20	120.12	59.70	41.98
Rat Clint (mL/min/kg)	55.20	20.77	41.80	59.40
Human Livermicrosome stability $t_{1/2}$ (min)	30.29	39.55	47.90	35.30
Human Clint (mL/min/kg)	52.90	40.53	33.50	45.50

F

parameter	16	30	31
Mouse $T_{1/2}$ (h) PO@20mg/kg	0.757	0.699	0.930
Mouse Tmax (h) PO@20mg/kg	0.667	0.417	0.333
Mouse Cmax (ng/mL) PO@20mg/kg	750.218	1095.919	1245.280
Mouse AUC(0-t) (h * ng/mL) PO@20mg/kg	1862.128	1556.770	1219.430
inh%CYP 2C9@10μM	4.33	NT	NT
inh% CYP 2D6@10μM	37.15	NT	NT
inh%CYP 3A4@10μM (Midazolam)	56.88	NT	NT
inh%CYP 3A4@10μM (Testosteron)	37.13	NT	NT
hERG inh%@30μM	38.94	NT	NT

notes:
NT: No testing; PO: peros

308
309 **Fig.5. Cellular activity and PK profiling.** A-C) Antiviral activity of compound **2**, **16**, **30**, **31** against SARS-
310 CoV-2 wild type (WT) (A), variants Omicron BA.5 (B), and XBB.1 (C). The EC_{50} was measured on virus
311 infected Vero E6 cells. The error bars are mean ± s.d based on three replicates. D) Toxicity to normal HEK293T
312 cell lines. Data are presented as mean of two technical replicates. E) Table for liver stability in rat and human
313 liver microsomes. F) Table for in vivo PK characteristics, CYP inhibition and hERG toxicity.

314

315 **Discussion**

316 In the current study, we leveraged an AI generative model to discover a novel series of PL^{pro}
317 inhibitors targeting the SARS-Cov-2 virus. Using the generative model, Tree-Invent, we
318 successfully identified a novel scaffold of PL^{pro} inhibitors exhibiting comparable activity to a
319 known reference compound **GZNL-P4**, though has much improved liver metabolic stability.
320 Further optimization work on this series results in a lead compound (**GZNL-2002**) with an
321 improved enzyme potency and cellular-level antiviral activity. This compound demonstrates
322 excellent liver stability, low CYP and hERG inhibition and it also demonstrate reasonable oral PK
323 properties. We believe this study offers valuable insights for the future development of PL^{pro}
324 inhibitors as a potential treatment for COVID-19 disease.

325

326 **Materials and Methods**

327 **General methods for chemical synthesis**

328 Synthesis routes: compound preparation and structure characterization were elaborately
329 stated in supplemental information. All compounds were isolated through column chromatography
330 or preparative liquid chromatography equipped with a PDA detector and a SB-C18 column (30 ×
331 150 mm) of waters. The mobile phase consisting of buffer A (ultrapure H_2O containing 0.1%
332 formic acid) and buffer B (chromatographic grade CH_3CN) was applied with a flow rate of 20
333 mL/min. All solvents and chemicals were used without further purification. ^1H and ^{13}C NMR
334 spectra were recorded on a Bruker Avance neo at 600 MHz for ^1H NMR and 100 MHZ for ^{13}C
335 NMR, respectively. Chemical shifts (δ) were expressed in parts per million (ppm) with TMS as an
336 internal standard. The letters s, d, t, and q represent multiplet. The purity of all tested compounds
337 was larger than 95%, as estimated using a Waters Acquity UPLC system equipped with a PDA
338 detector and a BEH-C18 column (2.1 × 50 mm). The mobile phase consists of buffer A (ultrapure
339 H_2O containing 0.1% formic acid) and buffer B (chromatographic grade CH_3CN) and the flow rate
340 is 0.4 mL/min.

341 **Molecular generation with the Tree-Invent model**

342 Tree-Invent is a topology constrained molecular generative model recently developed by our
343 group(*32*). The unique feature of the algorithm is utilizing a topological tree structure as a
344 constraint for molecular generation (Fig.6A). It is particularly suitable for lead optimization tasks,

such as scaffold hopping, R-group decoration etc, when only part of structure needs to be changed. Tree-Invent integrates the topological tree into an atom-based autoregressive structural generation process (For more details, please refer to (*32*)), where each node of the topological tree is sequentially decoded to either a single atom or a ring structure. Compared to available models specific for lead optimization, such as SAMOA(*42*), GraphGMVAE (*43*), and DeepHop (*44*), Tree-Invent provides a direct control on molecular generation.

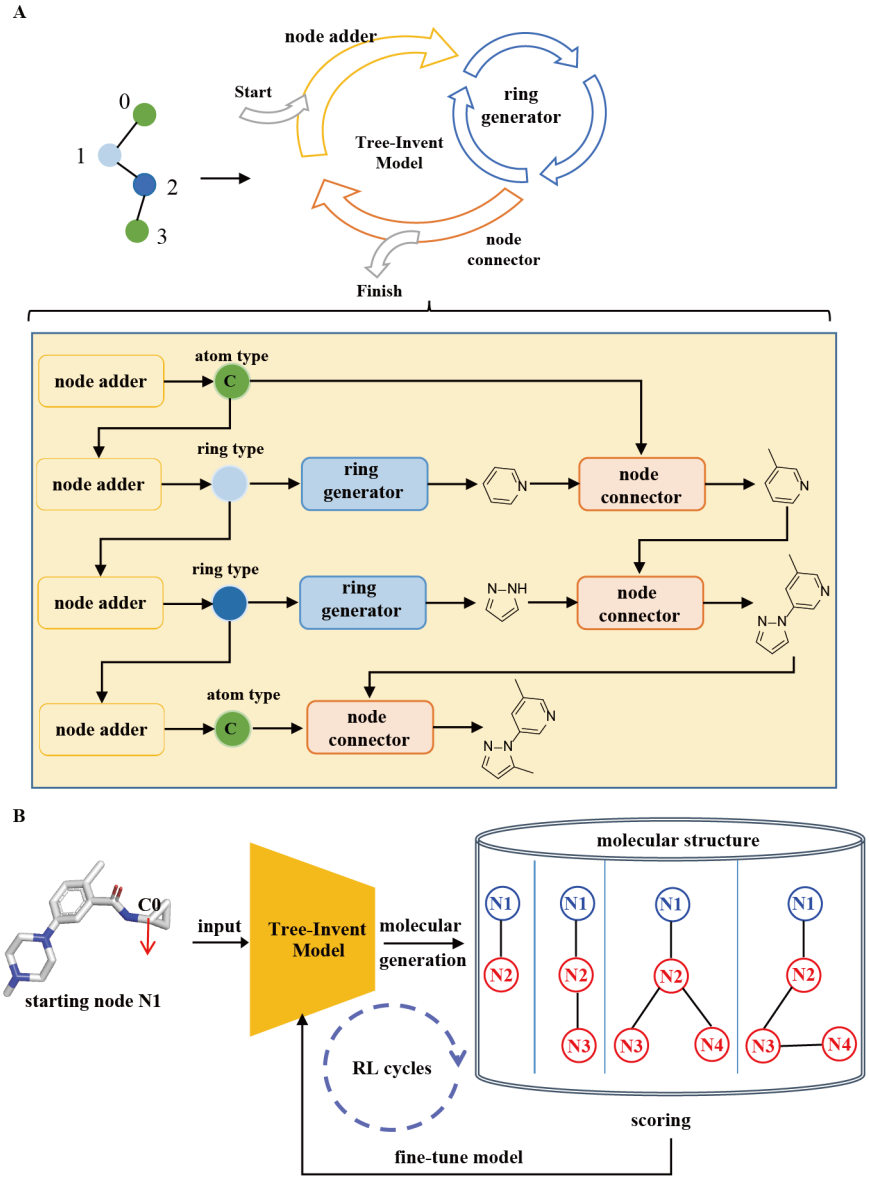


Fig.6. **Molecular generation method with Tree-Invent**. A) Scheme of molecular generation under a topology constraint. B) Basic workflow for molecular generation based on the substructure of GZNL-P4 using the Tree-Invent model with RL strategy.

The architecture of Tree-Invent contains three independent modules to complete molecular generation, namely node adder, ring generator and node connector (Fig.6A). The module of node adder determines the type of a node and whether the structure should be terminated at current status. The module of ring generator is executed only when a ring node is sampled. This module provides precise control on the generation of ring structure through constraining ring size, number of ring, and number of aromatic ring in the ring system. The module of node connector determines the bond type between the existing node and the new node and also the atom position in the existing node for node growth or connection. Particularly, the model can generate structures according to given topological constraints through masking the action distribution probability (ADP) (*32*). Through this unique masking mechanism, Tree-Invent can not only imposes precise control on generation of certain substructures but also allows generation of other parts of the structure in an uncontrolled manner. Furthermore, by incorporating reinforcement learning, Tree-Invent can achieve goal-directed structure generation by optimizing the reward for a given structure.

In this study, we transformed the 2D structure of a reference compound **GZNL-P4** into a simple topological tree (as shown in Fig.3B). One pre-defined node (i.e. node N1) represents the structure comprising the linker and tail parts of **GZNL-P4**. Among the atoms in the node, the atom C0 is regarded as the connection atom with new node N2 (Fig.6B), while other atoms are blocked to prevent connecting the new node by including them in the predefined saturation atom list (for configurating saturation atom list in node definition, please refer to supplemental material), which the connecting probability is masked to zero. These settings enable node N1 to be completely fixed in the process of molecular generation. Other parts of molecule including node N2, N3 and N4 are specified to only grow from node N1 sequentially, starting with a single bond, mimicking the head group of reference **GZNL-P4**. More specifically, node N2 is confined to have bicyclic ring systems, which contains at least one aromatic ring and at most two R-group substitutions. In the auto-regressive process of molecular generation, maximal node adding operation was set to four steps, in which the first two steps were specified to cover the pre-defined node N1 and bicyclic node N2. In the two additional steps (N3 and N4 node), adding substitutions on the bicyclic ring is done in uncontrolled mode (ie. no masking on the ADP) and their generation is purely dependent on the current state of the structure (Fig.6B), which means that the generation of these two nodes may be skipped by the model to generate naked ring as the head group. These topology constrains were pre-defined in the job configuration file.

388 In total, 1000 epochs of RL were run to generate molecules by rewarding molecules for better
389 docking score. The prior model for RL was trained on the GuacaMol dataset (*32*) and in the process
390 of RL, the learning rate was set to 0.00005 with the batch size of 20. The range of raw docking
391 score from -8 kcal/mol to -5 kcal/mol was used to define the Sigmoid transformation function of
392 docking score P ($P \in [0, 1]$), which is included in the scoring function of RL. DockStream(*45*), a
393 docking wrapper to facilitate *de novo* molecule design, was integrated in the Tree-Invent model to
394 integrate Glide docking using the high-throughput mode (Glide-HTVS)(*36*).
395
396 **Molecular docking**
397 Molecular docking was conducted utilizing the Glide docking program of Schrödinger
398 software suite (version 2020)[27]. The co-crystallized structure of the SARS-CoV-2 PL^{pro} enzyme in
399 complex with **GRL0617** (PDB code: 7JRN) was served as the protein model for the docking. The
400 protein structure underwent pre-processing using the Protein Preparation Wizard tool within the
401 software, ensuring proper assignment of bond orders and repair of amino acids with missing side
402 chains. Additionally, all crystal waters were removed. Subsequently, the Epik algorithm(*46*) was
403 employed to generate ionization and tautomeric states of the protein at a pH of 7.0. The structure
404 was then minimized using the OPLS3 force field(*47*) until heavy atoms converging to a root mean
405 square deviation of 0.3 Å from the original coordinates. The minimized structure was utilized to
406 prepare the receptor grid for Glide docking. Concurrently, ligands intended for docking were
407 prepared using the LigPrep module in Schrödinger. Analogous to protein preprocessing, ligands
408 were protonated using the Epik algorithm, and tautomeric and stereoisomeric forms were
409 enumerated. For each compound, a single low-energy conformation was generated utilizing the
410 OPLS3 force field. The prepared ligands were then docked into the predefined docking site within
411 a box size of 20 Å. To evaluate the generated molecules, the docking process employed standard
412 precision, allowing for flexible ligand conformation sampling. After docking, a maximal 10 poses
413 were retained for every ligand and ranked based on the Glidescore. Poses with a docking score
414 lower than -7.0 kcal/mol were retained. Besides, poses that did not fit to the binding mode of
415 **GRL0617** or compound **2**, particularly the stacking interaction with Y268 and P248/P247, as well
416 as hydrogen bonding with D164 and Q269 mediated by the amide group, were eliminated.
417
418 **Plasmid Construction, Protein Expression and Purification**

419 The papain-like protease domain sequence was retrieved from the complete genome of SARS-
420 CoV-2, accessible in the NCBI gene databank under the Gene ID: 43730578. Subsequently, the
421 protein sequence encoding the Ubl domain of the Nsp3 protein (spanning amino acids 746-1060,
422 accession number YP_009742610.1) was codon-optimized for enhanced expression in bacterial
423 systems. The optimized gene was then chemically synthesized and cloned into the pET28a
424 expression vector with BamHI and XhoI (Tsingke). A recognition sequence for the TEV protease
425 was strategically inserted between the N-terminal His6-tag and the PL^{pro} sequence to facilitate
426 subsequent tag removal.

427 The PL^{pro} mutant C111S (PL^{pro}C111S) was generated by site-directed mutagenesis using PCR.
428 For the protein expression, BL21(DE3) *Escherichia coli competent cells* were transformed with
429 the PL^{pro} expression plasmid. These cells were then inoculated into LB medium and grown to an
430 OD600 of 0.6-0.8. Subsequently, protein production was induced by adding 0.5 mM of isopropyl-
431 D-thiogalactopyranoside (IPTG) and 1mM zinc chloride ($ZnCl_2$) at 16℃ for 18 hours. After
432 induction, the cell pellets were collected through centrifugation at 4000 rpm for 30 minutes and
433 stored at −80 ℃ for future use. For protein purification, the frozen cell pellets were thawed and
434 resuspended in buffer A, containing 50 mM Tris-HCl, 150 mM NaCl, 10 mM imidazole, and 1
435 mM TCEP (pH 7.4), supplemented with lysozyme, deoxyribonuclease, and a protease inhibitor
436 cocktail (Roche) to protect the target protein from degradation. Cell lysis was achieved through
437 sonication, and the lysates were separated from cellular debris by centrifugation at 18,000 × g for
438 35 minutes, repeated twice at 4 ℃. The clarified supernatant was further filtered using a 0.22 μm
439 pore size membrane to remove any remaining particulate matter. Subsequently, the supernatant
440 containing His-tagged target protein was loaded onto a HisTrap HP column (5 ml; GE) pre-
441 equilibrated with buffer A. After removal of nonspecific binding proteins by washing with buffer
442 A, the target protein was eluted with a linear gradient of buffer B, containing 50 mM Tris-HCl,
443 150 mM NaCl, 250 mM imidazole, and 1 mM TCEP (pH 7.4) over 20 column volumes.

444 To remove the His-tag, the eluted protein was dialyzed in the presence of TEV protease.
445 Following the removal of His-tag, the protein solution was concentrated to 2 ml and loaded onto a
446 Superdex 75 column (120 ml; GE) that had been pre-equilibrated with buffer C, containing 20 mM
447 Tris-HCl, 100 mM NaCl, and 1 mM TCEP (pH 7.4). This final chromatographic step was
448 performed to achieve high purity and homogeneity of the target protein, ready for downstream
449 applications.

450 **PL^{pro} enzymatic assay**

451 The activity of PL^{pro} was assessed using the labeled peptide Z-RLRGG-AMC (GLPBIO,
452 GA23715) as a substrate in a 384-well plate (GREINER, 784076), with the excitation and emission
453 wavelengths at 340 nm and 460 nm. The enzymatic reaction was carried out in the buffer
454 containing 50mM HEPES, 10mM DTT, 0.1mM EDTA, 0.005% tween 20, pH7.2, with GRL0617
455 (TargetMol, Shanghai, China) as the positive control. The serially diluted compounds were added
456 into PL^{pro} solution diluted with assay buffer and incubated for 10 minutes. Then, the diluted
457 substrate peptide with assay buffer was added into PL^{pro} and inhibitor mix solution and incubated
458 for 1 hour at 37 °C. The final concentrations of components in the reaction system are 10 nM and
459 20 μM for PL^{pro} and substrate, respectively. Fluorescence intensity was recorded with a BioTek
460 Neo2 multimode plate reader (Agilent). The inhibition activity IC_{50} values were calculated using
461 nonlinear regression fitting in GraphPad Prism 9.0.

462 **Differential scanning fluorimetry assay**

463 The fluorescence dye SYPRO Orange (Sigam-Aldrich, S5692) was used as a fluorescence
464 probe in differential scanning fluorimetry assay. The protein sample was diluted to 10μM in PBS
465 (pH7.4) and incubated with 20 μM of compounds, with DMSO as control. The fluorescence dye
466 SYPRO Orange (Sigam-Aldrich, S5692) was then added to the protein samples at the 5x dye
467 concentration. The sample was mixed thoroughly and transferred to the 384-well microplate (10
468 μL per well). Protein buffer containing DMSO was used as a blank control. The microplate was
469 placed in the fluorescence quantitative PCR instrument. Fluorescence was monitored when the
470 temperature was gradually raised from 30 to 80 °C in 0.2 °C increments at 5-second intervals.
471 Three technical replicates were tested for each sample. Melt curve data were plotted using the
472 Boltzmann model to obtain the melt temperature (Tm, the midpoint of unfolding of the protein)
473 using GraphPad Prism 9.0.

474 **Crystallization, data collection and structure determination**

475 The purified SARS-CoV-2 PL^{pro}C111S protein was incubated overnight with compound **2** at
476 a 1:3 protein-to-ligand molar ratio in buffer C, and concentrated to 8 mg/ml. Crystal screening was
477 carried out using the sitting-drop vapor diffusion technique at 4 °C in 96-well crystallization plates.
478 Each crystallization drop contained a 1:1 mixture of the protein complex sample and reservoir
479 solution. The protein-ligand complex crystals were obtained after 7-10 days after the droplet
480 setting. Optimal crystallization conditions were determined for compound **2**: 0.05 M Zinc acetate

481 dihydrate, 22% w/v Polyethylene glycol 3,350. X-ray diffraction data were collected at the
482 Shanghai Synchrotron Radiation Facility's beamlines BL02U1. Data processing was performed
483 using the Aquarium program (*48*). The protein structures were solved by molecular replacement
484 using the program Phaser (*49*) integrated in PHENIX package (*50*), employing the previously
485 published PL^{pro} structure (PDB code: 7CJD)(*51*) as a search model. Structure refinement was
486 carried out in PHENIX, with manual interactive model rebuilding conducted in Coot (*52*). Ligand
487 constraints were generated using Phenix.elbow (*53*), and ligand models were built based on the
488 omit map. The data collection and refinement statistics are summarized in Table S1.

489 **Cell culture and cell viability assay**

490 Vero E6 (ATCC, CRL-1586) and HEK293T (ATCC CRL-3216) cells were maintained in
491 Dulbecco's Modified Eagle Medium (DMEM) enriched with 10% fetal bovine serum (FBS), 100
492 IU/mL penicillin, and 100 μg/mL streptomycin. Cells were incubated at 37 °C in a humidified
493 atmosphere of 5% CO_2 and confirmed to be free of mycoplasma contamination. Cell viability was
494 assessed using the Cell Titer-Glo 2.0 Cell Viability Assay (G9242, Promega) following the
495 manufacturer's protocol. Briefly, 4×10^3 cells suspended in 20 μL of culture medium were plated
496 in opaque-walled 384-well plates and allowed to adhere for 12 hours. After a 48-hour treatment
497 with compounds, 20 μL of Cell Titer-Glo reagent was dispensed into each well. Following 2
498 minutes of shaking and 10 minutes of incubation, luminescence was quantified using a BioTek
499 Synergy H1 multimode reader (Agilent).

500 **Virus preparation and titrations**

501 SARS-CoV-2 wild-type strain (WT, GenBank: MT123291), the Omicron BA.5 variant
502 (Omicron BA.5, GDPCC-303), and the XBB.1 variant (XBB.1, IQTC-1596943) were propagated
503 in Vero E6 cells using established protocols (*54, 55*) and stored at -80 °C for future use. Virus
504 titers were determined by performing 10-fold serial dilutions in confluent Vero E6 cells seeded in
505 96-well microtitre plates. After three days of post-inoculation, a cytopathic effect (CPE) was
506 observed and scored, and the Reed-Muench formula was employed to calculate the TCID50,
507 representing the tissue culture infectious dose causing CPE in 50% of the cell cultures. All
508 infection experiments were conducted within a Biosafety Level 3 (BSL-3) facility at the
509 Guangzhou Customs Inspection and Quarantine Technology Center (IQTC), adhering to strict
510 safety protocols to ensure the safety of personnel and the environment.

511 ***In vitro* antiviral activity assay**

512 The in vitro antiviral activity was performed in Vero E6 cell lines using the CPE method (*54*).
513 A 96-well plate was seeded with 2×10^4 Vero E6 cells and incubated for 24 hours to allow for cell
514 attachment and monolayer formation. Subsequently, various dilutions of the compounds were
515 combined with SARS-CoV-2 at a multiplicity of infection (MOI) of 0.01. Each well received 200
516 μL of this compound-virus mixture, which was inoculated onto the monolayer of Vero E6 cells.
517 Seventy-two hours post-inoculation, the CPE was quantified using a Celigo Image Cytometer. The
518 inhibitory activity of the compounds was determined based on the CPE rates observed for SARS-
519 CoV-2. The EC_{50} value, representing the effective concentration at which 50% of the CPE is
520 inhibited, was calculated from these data. Each experiment was conducted independently three
521 times, with eight concentration gradients tested in triplicate wells. Representative data from one
522 of these experiments are presented.

523 **Metabolic stability in liver microsome**

524 To assess the in vitro metabolic stability, a series of incubations were conducted using rat and
525 human liver microsomes. Each compound was diluted to a concentration of 1.5 μM and incubated
526 with either rat or human liver microsomes (0.75 mg/mL) in a 0.5 mL volume of PBS. Assay plates
527 were prepared with 30 μL aliquots of the compound-microsome mixture, designated for various
528 time points (0, 5, 15, 30, and 45 minutes). The reaction was initiated by adding NADPH to a final
529 concentration of 2 mM, and the mixtures were incubated at 37 °C. At each designated time point,
530 the reaction was terminated by adding an acetonitrile solution containing internal standard (IS,
531 Tolbuamide). The plates were then shaken at 600 rpm for 10 minutes followed by centrifugation
532 at 6000 rpm for 15 minutes. This centrifugation step separated the precipitated protein from the
533 supernatant, which was then transferred to a 96-well plate containing 140 μL of pure water. The
534 samples were analyzed using LC/MS (Shimadzu Nexera LC-40 & SCIEX TQ-6500+) to quantify
535 the remaining concentrations of the compounds. The intrinsic clearance, representing the rate of
536 substrate depletion, was calculated by considering the amount of microsomes per liver weight and
537 the liver weight per kilogram of body weight.

538 **In vivo PK study**

539 In this study, pharmacokinetic assessments were conducted on SPF male ICR mice (n = 3) at
540 Medicilon (Shanghai, China). The mice were dosed with 20mg/kg of the test compound using an
541 oral administration route. The dosing vehicle comprised a solution containing 5% DMSO, 10% of
542 the test compound, and 85% saline. Sequential blood samples were collected from the mice at

543 various time points post-administration: 0.25h, 0.5h, 1h, 2h, 4h, 6h, 8h, and 24h. These samples
544 were then analyzed using liquid chromatography tandem mass spectrometry (LC-MS/MS) to
545 quantify the concentration of the compound in the blood. To assess the pharmacokinetic behavior
546 of the compound, key parameters were calculated using Phoenix WinNonlin 7.0 software. These
547 parameters included the area under the concentration-time curve from time zero to the last
548 measurable concentration (AUC_{0-t}), the maximum plasma concentration (C_{max}), the time to reach
549 Cmax (T_{max}), and the elimination half-life ($T_{1/2}$).

550 **CYP inhibition**

551 The compound was added into the liver microsomes solution pre-warmed at 37 °C for 5 min.
552 Subsequently, 20 μL of a pre-warmed (37 °C) 10 mM NADPH solution was introduced into 180
553 μL of the liver microsomes solution containing the compounds, thus initiating the enzyme reaction.
554 The assay plate was then incubated at 37 °C until the designated time point, at which 300μL of a
555 termination solution containing 35 ng/mL of ketoprofen, 7.5 ng/mL of carbamazepine, 5 ng/mL of
556 diphenhydramine, and 10 ng/mL of tolbutamide was added to halt the reaction. Following
557 termination, the samples were centrifuged at 3,220 × g for 40 minutes. The supernatant was
558 carefully transferred to an analysis plate containing an appropriate volume of ultra-pure water,
559 ready for LC-MS/MS analysis. Through this process, the inhibition rate was determined for the
560 CYP isoforms 2C9, 2D6, and 3A4.

561 **hERG inhibition**

562 A patch clamp system procured from Sutter Instrument (USA) was employed to monitor the
563 current in human embryonic kidney cells stably expressing hERG channels (hERG-HEK293 cells,
564 Creacell, A-0320). The cells were clamped using the patch clamp system and a whole-cell voltage-
565 clamp mode was executed to elicit hERG currents at corresponding voltages. Both cisapride (a
566 positive control from sigma) and test compounds were applied to the cells in serially diluted
567 concentrations. Tail currents of the hERG channels were meticulously recorded using Patchcontrol
568 HT and Patchmaster to obtain the peak tail current at each concentration. The current recorded in
569 a non-compound containing solution served as a control for each cell. For each cell, the current
570 was recorded independently twice, and at least two cells were recorded for each concentration.

571 **Data analysis**

572 The data in the figures represent mean ± SD. All data were analyzed using GraphPad Prism
573 9.0 software. Statistical comparison between different groups was performed using corresponding

statistical analysis labeled in figure legends combining several experiments. The binding mode of ligands with protein was illustrated by PyMOL2.6.0a0 (open source).

References

1. M. Li *et al.*, COVID-19 vaccine development: milestones, lessons and prospects. *Signal transduction and targeted therapy* **7**, 146 (2022).
2. W. T. Harvey *et al.*, SARS-CoV-2 variants, spike mutations and immune escape. *Nature reviews. Microbiology* **19**, 409-424 (2021).
3. G. Li, R. Hilgenfeld, R. Whitley, E. De Clercq, Therapeutic strategies for COVID-19: progress and lessons learned. *Nature reviews. Drug discovery* **22**, 449-475 (2023).
4. A. Amstutz *et al.*, Effects of remdesivir in patients hospitalised with COVID-19: a systematic review and individual patient data meta-analysis of randomised controlled trials. *The Lancet. Respiratory medicine* **11**, 453-464 (2023).
5. R. L. Gottlieb *et al.*, Early remdesivir to prevent progression to severe Covid-19 in outpatients. *The New England journal of medicine* **386**, 305-315 (2022).
6. M. Yasir *et al.*, An updated systematic review on Remdesivir's safety and efficacy in patients afflicted with COVID-19. *Cureus* **15**, e43060 (2023).
7. F. Kabinger *et al.*, Mechanism of molnupiravir-induced SARS-CoV-2 mutagenesis. *Nature structural & molecular biology* **28**, 740-746 (2021).
8. M. D. Waters, S. Warren, C. Hughes, P. Lewis, F. Zhang, Human genetic risk of treatment with antiviral nucleoside analog drugs that induce lethal mutagenesis: The special case of molnupiravir. *Environmental and molecular mutagenesis* **63**, 37-63 (2022).
9. D. R. Owen *et al.*, An oral SARS-CoV-2 M(pro) inhibitor clinical candidate for the treatment of COVID-19. *Science (New York, N.Y.)* **374**, 1586-1593 (2021).
10. Y. Unoh *et al.*, Discovery of S-217622, a noncovalent oral SARS-CoV-2 3CL protease inhibitor clinical candidate for treating COVID-19. *Journal of medicinal chemistry* **65**, 6499-6512 (2022).
11. R. Shimizu *et al.*, Evaluation of drug-drug interactions of Ensitrelvir, a SARS-CoV-2 3CL protease inhibitor, with transporter substrates based on in vitro and clinical studies. *Journal of clinical pharmacology* **63**, 918-927 (2023).
12. S. Iketani *et al.*, Multiple pathways for SARS-CoV-2 resistance to nirmatrelvir. *Nature* **613**, 558-564 (2023).
13. S. Gandhi *et al.*, De novo emergence of a remdesivir resistance mutation during treatment of persistent SARS-CoV-2 infection in an immunocompromised patient: a case report. *Nature communications* **13**, 1547 (2022).
14. B. H. Harcourt *et al.*, Identification of severe acute respiratory syndrome coronavirus replicase products and characterization of papain-like protease activity. *Journal of virology* **78**, 13600-13612 (2004).
15. D. Shin *et al.*, Papain-like protease regulates SARS-CoV-2 viral spread and innate immunity. *Nature* **587**, 657-662 (2020).
16. S. Kandwal, D. Fayne, Genetic conservation across SARS-CoV-2 non-structural proteins - Insights into possible targets for treatment of future viral outbreaks. *Virology* **581**, 97-115 (2023).

17. J. Osipiuk *et al.*, Structure of papain-like protease from SARS-CoV-2 and its complexes with non-covalent inhibitors. *Nature communications* **12**, 743 (2021).
18. C. Ma *et al.*, Discovery of SARS-CoV-2 papain-like protease inhibitors through a combination of high-throughput screening and a FlipGFP-based reporter assay. *ACS central science* **7**, 1245-1260 (2021).
19. Z. Shen *et al.*, Design of SARS-CoV-2 PLpro inhibitors for COVID-19 antiviral therapy leveraging binding cooperativity. *Journal of medicinal chemistry* **65**, 2940-2955 (2022).
20. P. Jadhav *et al.*, Structure-based design of SARS-CoV-2 papain-like protease inhibitors. *European journal of medicinal chemistry* **264**, 116011 (2024).
21. B. C. Sanders *et al.*, Potent and selective covalent inhibition of the papain-like protease from SARS-CoV-2. *Nature communications* **14**, 1733 (2023).
22. B. Tan *et al.*, Design of a SARS-CoV-2 papain-like protease inhibitor with antiviral efficacy in a mouse model. *Science (New York, N.Y.)* **383**, 1434-1440 (2024).
23. Y. Lu *et al.*, Discovery of orally bioavailable SARS-CoV-2 papain-like protease inhibitor as a potential treatment for COVID-19. *bioRxiv*, 2024.2004.2003.587743 (2024).
24. W. He *et al.* (China, 2022).
25. V. Chenthamarakshan *et al.*, Accelerating drug target inhibitor discovery with a deep generative foundation model. *Science advances* **9**, eadg7865 (2023).
26. R. Gómez-Bombarelli *et al.*, Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science* **4**, 268-276 (2018).
27. Y. Yang *et al.*, SyntaLinker: automatic fragment linking with deep conditional transformer neural networks. *Chemical science* **11**, 8312-8322 (2020).
28. S. Kang, K. Cho, Conditional molecular design with deep generative models. *Journal of chemical information and modeling* **59**, 43-52 (2019).
29. S. Amabilino, P. Pogány, S. D. Pickett, D. V. S. Green, Guidelines for recurrent neural network transfer learning-based molecular generation of focused libraries. *Journal of chemical information and modeling* **60**, 5699-5713 (2020).
30. M. Olivecrona, T. Blaschke, O. Engkvist, H. Chen, Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics* **9**, 48 (2017).
31. O. Zhang *et al.*, Learning on topological surface and geometric structure for 3D molecular generation. *Nature computational science* **3**, 849-859 (2023).
32. M. Xu, H. Chen, Tree-Invent: A novel multipurpose molecular generative model constrained with a topological tree. *Journal of chemical information and modeling* **63**, 7067-7082 (2023).
33. Y. Matsuo *et al.*, Deep learning, reinforcement learning, and world models. *Neural networks : the official journal of the International Neural Network Society* **152**, 267-275 (2022).
34. C. Y. Chou *et al.*, Structural basis for catalysis and ubiquitin recognition by the severe acute respiratory syndrome coronavirus papain-like protease. *Acta crystallographica. Section D, Biological crystallography* **70**, 572-581 (2014).
35. H. Cho *et al.*, In vitro metabolic characterization of the SARS-CoV-2 papain-like protease inhibitors GRL0617 and HY-17542. *Frontiers in pharmacology* **14**, 1067408 (2023).
36. R. A. Friesner *et al.*, Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of medicinal chemistry* **47**, 1739-1749 (2004).

37. C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews* **46**, 3-26 (2001).
38. M. Ringnér, What is principal component analysis? *Nature biotechnology* **26**, 303-304 (2008).
39. H. Zhou, J. Skolnick, Utility of the morgan fingerprint in structure-based virtual ligand screening. *The journal of physical chemistry. B* **128**, 5363-5370 (2024).
40. G. C. Linderman, S. Steinerberger, Clustering with t-SNE, provably. *SIAM journal on mathematics of data science* **1**, 313-332 (2019).
41. K. Ratia *et al.*, A noncovalent class of papain-like protease/deubiquitinase inhibitors blocks SARS virus replication. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 16119-16124 (2008).
42. M. Langevin, H. Minoux, M. Levesque, M. Bianciotto, Scaffold-constrained molecular generation. *Journal of chemical information and modeling* **60**, 5637-5646 (2020).
43. Y. Yu *et al.*, A novel scalarized scaffold hopping algorithm with graph-based variational autoencoder for discovery of JAK1 inhibitors. *ACS omega* **6**, 22945-22954 (2021).
44. S. Zheng *et al.*, Deep scaffold hopping with multimodal transformer neural networks. *Journal of cheminformatics* **13**, 87 (2021).
45. J. Guo *et al.*, DockStream: a docking wrapper to enhance de novo molecular design. *Journal of cheminformatics* **13**, 89 (2021).
46. J. C. Shelley *et al.*, Epik: a software program for pK(a) prediction and protonation state generation for drug-like molecules. *Journal of computer-aided molecular design* **21**, 681-691 (2007).
47. E. Harder *et al.*, OPLS3: A force field providing broad coverage of drug-like small molecules and proteins. *Journal of chemical theory and computation* **12**, 281-296 (2016).
48. F. Yu *et al.*, Aquarium: an automatic data-processing and experiment information management system for biological macromolecular crystallography beamlines. *Journal of Applied Crystallography* **52**, 472-477 (2019).
49. A. J. McCoy *et al.*, Phaser crystallographic software. *J Appl Crystallogr* **40**, 658-674 (2007).
50. P. D. Adams *et al.*, The Phenix software for automated determination of macromolecular structures. *Methods (San Diego, Calif.)* **55**, 94-106 (2011).
51. X. Gao *et al.*, Crystal structure of SARS-CoV-2 papain-like protease. *Acta pharmaceutica Sinica. B* **11**, 237-245 (2021).
52. P. Emsley, K. Cowtan, Coot: model-building tools for molecular graphics. *Acta crystallographica. Section D, Biological crystallography* **60**, 2126-2132 (2004).
53. N. W. Moriarty, R. W. Grosse-Kunstleve, P. D. Adams, electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation. *Acta crystallographica. Section D, Biological crystallography* **65**, 1074-1080 (2009).
54. Y. Guo *et al.*, Discovery and characterization of potent pan-variant SARS-CoV-2 neutralizing antibodies from individuals with Omicron breakthrough infection. *Nature communications* **14**, 3537 (2023).
55. H. Wang *et al.*, Structure-based discovery of dual pathway inhibitors for SARS-CoV-2 entry. *Nature communications* **14**, 7574 (2023).

Acknowledgments

We are grateful for support from Shanghai Synchrotron Radiation Facility (SSRF) beamlines (BL19U1 and BL02U1) and Guangzhou Laboratory Core Facilities.

Funding:

Overseas Experts Supporting Programs under National Research Platform (WGZJ22-001); National Natural Science Foundation of China (82170473); The startup and R&D Program of Guangzhou National Laboratory (GZNL2024A01005, GZNL2023A01008, GZNL2023A02012, GZNL2024A01022, YW-YWYM0205, YW-YWYM0202, SRPG22-002, SRPG22-003, SRPG22-011); Guangdong Natural Science Foundation (2021QN020451, 2021CX020227, 2024A1515011589); Basic and Applied Basic Research Projects of Guangzhou Basic Research Program (202201011795, 2023A04J0161).

Author contributions:

Writing: J.S., T.R., Y.L. and P.Z.

Medicinal chemistry: T.R., G.Z., P.Z., H.C., P.H. and J.Z.

Computational studies: T.R., M.X., M.T., and C.H.

Cellular and biochemical assays: Y.L., Q.Y., W.L., M.D., J.C., J.Z., A.Z., K.W. and X.W.

Crystallography: Y.L. and J.S. performed crystallography.

Cloning and/or purified proteins: Y.L., W.L., M.D., J.C., J.Z., K.W. and X.W.

Supervision: J.S., C.X., and C.H.

Competing interests:

The authors declare no competing interests.

Data and materials availability:

Crystal structures generated during the current study are available in the Protein Data Bank (PDB) under accession codes 8ZSE (PL^{pro} bound to compound 2). All other datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.