Unveiling the molecular mechanism of SARS-CoV-2 main protease inhibition from 137 crystal structures using algebraic topology and deep learning†

Duc Duy Nguyen,a Kaifu Gao,b Jiahui Chen,b Rui Wangb and Guo-Wei Wei†abc

Currently, there is neither effective antiviral drugs nor vaccine for coronavirus disease 2019 (COVID-19) caused by acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Due to its high conservativeness and low similarity with human genes, SARS-CoV-2 main protease (M\textsuperscript{pro}) is one of the most favorable drug targets. However, the current understanding of the molecular mechanism of M\textsuperscript{pro} inhibition is limited by the lack of reliable binding affinity ranking and prediction of existing structures of M\textsuperscript{pro}-inhibitor complexes. This work integrates mathematics (i.e., algebraic topology) and deep learning (MathDL) to provide a reliable ranking of the binding affinities of 137 SARS-CoV-2 M\textsuperscript{pro} inhibitor structures. We reveal that Gly143 residue in M\textsuperscript{pro} is the most attractive site to form hydrogen bonds, followed by Glu166, Cys145, and His163. We also identify 71 targeted covalent bonding inhibitors. MathDL was validated on the PDBbind v2016 core set benchmark and a carefully curated SARS-CoV-2 inhibitor dataset to ensure the reliability of the present binding affinity prediction. The present binding affinity ranking, interaction analysis, and fragment decomposition offer a foundation for future drug discovery efforts.

1 Introduction

Starting in late Dec, 2019, the COVID-19 pandemic caused by new severe acute respiratory syndrome coronavirus (SARS-CoV-2) has infected more than 22 million individuals and has caused more than 777,000 fatalities in all of the continents and over 213 countries and territories by August 19th, 2020. Under the current global health emergency, researchers around the world have engaged in the investigation of the different drug targets of SARS-CoV-2, such as the main protease (M\textsuperscript{pro}, also called 3CL\textsuperscript{pro}), papain-Like protease (PL\textsuperscript{pro}), RNA-dependent RNA polymerase (RdRp), 5'-to-3' helicase protein (Nsp13) to seek potential cures for this serious pandemic. To date, although there are some vaccines undergoing the Phase III trials, their safety and efficacy are still unclear.²

The main protease, one of the best-characterized targets for coronaviruses, attracts lots of research attention because it is very conservative and distinguished from any human gene. A recent study shows that although the overall sequence identity between SARS-CoV and SARS-CoV-2 is just 80%, the M\textsuperscript{pro} of SARS-CoV-2 shares 96.08% sequence identity to that of SARS-CoV.³ Therefore, we hypothesize that a potent SARS M\textsuperscript{pro} inhibitor is also a potent SARA-CoV-2 M\textsuperscript{pro} inhibitor.

At this moment, more than 300 potential SARS CoV M\textsuperscript{pro} inhibitors with its binding affinities are available in ChEMBL database⁴ which can be considered as the potential SARS-CoV-2 M\textsuperscript{pro} inhibitors. Recently, total 146 crystal structures of SARS-CoV-2 M\textsuperscript{pro} with its ligand complexes are released on the Protein Data Bank (PDB).⁵ Among them, 137 crystal structures have no available binding affinities reported for various reasons. However, the central dogma of drug design and discovery concerns the molecular mechanism and binding affinity of drug target interactions. Knowing the binding affinities and their ranking of 137 SARS-CoV-2 M\textsuperscript{pro} inhibitors is of great significance to the future design of anti SARS-CoV-2 drugs.

In this work, for the first time, we predict the binding affinities of these 137 M\textsuperscript{pro}-inhibitor complexes by reformulating algebraic topology-based mathematics-deep learning (MathDL) models, which have been the top competitor in D3R Grand Challenges, a worldwide competition series in computer aided drug design in the past three years.⁶ We generate reliable poses for 141 M\textsuperscript{pro} inhibitors with binding affinities but without complex structures. Together with 44 other complexes, we compose a set of 185 M\textsuperscript{pro}-inhibitor complexes, which is paired with 17,382 protein–ligand complexes in PDBbind 2019 general set. These datasets are utilized to construct 11 MathDL models.
in single-task and multitask settings. One of these 11 MathDL models has been validated by using the PDBbind v2016 core set benchmark, achieving the top performance over all exiting scoring functions. The other ten MathDL models have cross-validated on a set of 185 M<sup>pro</sup>–inhibitor complexes, showing an averaged Pearson’s correlation coefficient of 0.73.

Notably, for covalent inhibitors, the scheme of covalent irreversible inhibition of SARS-CoV/SARS-CoV-2 M<sup>pro</sup> is presented below:

\[ E + I \overset{k_1}{\rightarrow} EI \overset{k_2}{\rightarrow} E - I \]

The inhibitor first binds to the protease noncovalently, then a nucleophilic attacking by Cys145 leads to the formation of a stable covalent bond between the protease and the inhibitor.\(^7\)\(^,\)\(^8\)\(^n\)

The interaction depends on both the equilibrium-binding constant \(k_1\) (designated as \(k_1/k_2\)) and the inactivation rate constant for covalent bond formation \(k_2\). In this work, the binding affinity/IC<sub>50</sub> assesses the first step to form noncovalent binding.

In a nutshell, the present work provides reliable binding affinity predictions and ranking of 137 SARS-CoV-2 inhibitors that have crystal structures. It also offers data curation and validated models for exploring potential SARS-CoV-2 M<sup>pro</sup> inhibitors. Furthermore, this work explores different possible binding regions on the SARS-CoV-2 main protease and decode the most favorable molecular fragments for the inhibitor design.

## 2 Results and discussions

### 2.1 Results

This section is devoted to the utilization of our MathDL models developed in Section 3.3 to predict the binding affinities and their ranking of SARS-CoV-2 inhibitors that do not have reported experimental affinities. To reduce the role of 3D pose prediction errors in our model, we use the SARS-CoV-2 inhibitors with X-ray structures available in the PDB for our study. We manually search these ligands on the PDB and arrive at a set consisting of 137 SARS-CoV-2 M<sup>pro</sup> inhibitors having X-ray crystal structures but lacking of experimental binding affinities. We name this set SARS-CoV PDB-noBA (see Table 3). In this experiment, we develop a MathDL model optimized from PDBbind v2016 core set (see Section 3.3.1), five MathDL-ALL and five MathDL-MT models obtained from 5-fold study on the SARS-CoV BA set (see Section 3.3.2). The final predicted binding affinity is the consensus of these 11 models. The top ten inhibitors indicated by our models are shown in Table 1.

The most potent SARS-CoV-2 inhibitor found by our MathDL models is the inhibitor Nol in complex 7c8t. Nol was synthesized by Yang and his colleagues,\(^*\) Nol is found remarkable activities against SARS-CoV and HCoV.\(^*\) Specifically, the dissociation constant \(K_i\) of Nol was found to be 0.053 μM against SARS-CoV.\(^*\) Our MathDL reveals that Nol still inhibits SARS-CoV-2 main protease with a potent affinity at 0.30 μM.

Another important top potent SARS-CoV-2 inhibitor found by our models is the Michael acceptor inhibitor N3 in complex 7bqy. Designed by Yang and his colleagues,\(^*\) N3 was found to have viral activities against different coronavirus M<sup>pro</sup> such as SARS-CoV and MERS-CoV,\(^*\)\(^*\)\(^n\) Specifically, the dissociation constant \(K_i\) of N3 was found to be 9.0 μM against SARS-CoV.\(^*\) Our MathDL reveals that N3 still inhibits SARS-CoV-2 main protease with an even better affinity at 0.59 μM. This finding is consistent with the literature work\(^*\) showing that N3 is a potent inhibitor of COVID-19 virus M<sup>pro</sup>.

The inhibitor Qys in the complex 6xmk is also noticeable. Our predicted IC<sub>50</sub> is 0.78 μM. Soon after we made the prediction, on August 12th, 2020, Rathnayake et al.\(^*\) released another Qys-main protease complex with PDB ID 6w2a and also reported the IC<sub>50</sub> of Qys to SARS-CoV-2 is 0.45 μM, which is close to our prediction.

It is worth pointing out, except for the inhibitor T9j in the complex 5rg1, the rest of inhibitors reported in Table 1 are covalent inhibitors, which irreversibly form covalent bonds with Cys145 of the main protease (see discussion in Section 2.2.2). However, our models only predict the non-covalent binding affinity which is measured before the enzyme deactivation. The predicted binding affinities of all 137 complexes in SARS-CoV PDB-noBA dataset from various MathDL models are presented in Table S8 in ESI.\(^*\) In this table, we also supply the synthetic accessibility score (SAS), partition coefficient log P, and solubility log S for each small molecule. Except for SAS obtained via RDKit,\(^*\)\(^1\)\(^3\) log P and log S are evaluated by our TopP-S model.\(^*\)

### 2.2 Discussion

#### 2.2.1 Binding site analysis.

Based on the crystal structure information of 137 complexes in SARS-CoV PDB-noBA set, we have identified 13 distinct binding site regions of the SARS-CoV-2 main protease as illustrated in Fig. 1. Those binding pockets are denoted by P<sub>i</sub>, \(i = 1, 2, \ldots, 13\). Fig. 2a reveals that binding pocket P<sub>1</sub> is the most common binding region of the SARS-CoV-2 main protease, which attracts around 80.2% of ligands in the SARS-CoV PDB-noBA data set of 137 complexes. This finding is no surprise since the binding pocket P<sub>1</sub> shares similar active sites to its predecessor, \(i.e.,\) SARS-CoV M<sup>pro</sup>. Specifically, P<sub>1</sub> encompasses His141 and Cys145 catalytic dyads which are imperative to the substrate-binding mechanism.\(^*\) In additions,
the substrate-binding residues Tyr161 and His163 (ref. 15) are covered in P1. Binding pockets P2, P3, P5, P7, P8, and P10 are the least favor sites consisting of only one ligand. The rest of the binding pockets involve no more than 7 ligands. To study the correlation of the binding regions to the binding free energy, we present the box plot in Fig. 2b to illustrate the energy values through their quartiles.

The prevailing binding pocket P1 is the best region on the SARS-CoV-2 Mpro for inhibitor design with the median binding energy being $-7.22 \text{ kcal mol}^{-1}$. Nol is the best inhibitor candidate for the binding site P1 with predicted affinity found to be $-8.90 \text{ kcal mol}^{-1}$. Other binding regions such as P4, and P11 are less common but show their adequate effects on the binding mechanism with their best energy binding affinities calculated at $-7.28 \text{ kcal mol}^{-1}$ and $-6.80 \text{ kcal mol}^{-1}$, respectively. These potential binding sites can guide drug combination to inhibit coronavirus Mpro effectively.

### 2.2.2 Interaction analysis.

By looking further into the interactions between the top inhibitors and the main protease, we have found that Nol, V2m, N3 are peptidomimetic inhibitors, they form as many as 8, 8, 9 hydrogen bonds respectively to the nearby residues and also all form 1 covalent bond with Cys145 as listed in Table 2 and depicted in Fig. 3. All these hydrogen bonds justify their potency of the first step of non-covalent binding to the main protease complex and confirms the robustness of our MathDL models; the covalent bonds make the binding irreversible. We also notice that these three inhibitors share two common hydrogen bonds to His163, His164 (see Table 2, Fig. 3a, c and d). Therefore, they have some similar predicted binding energies, especially 6xhm and 7bqy at $-8.50 \text{ kcal mol}^{-1}$ and $-8.49 \text{ kcal mol}^{-1}$, respectively. This examination manifests how well our models preserve and capture the physical and chemical properties described in intermolecular bonding interactions. Furthermore, the ligand T9J that binds to Mpro in complex 5rg1 with a quite close binding energy at $-8.50 \text{ kcal mol}^{-1}$ forms different hydrogen bonds in comparison to three previously mentioned inhibitors (see Table 2). Since our models only concern the non-covalent binding affinity, the lack of covalent bond in 5rg1’s interactions does not downgrade its binding strength. With two relatively large hydrogen bonding distances (O2-His163: 3.05 Å, O3-Glu166: 3.38 Å (see Fig. 3d)), the binding affinity of 5rg1 is still comparable to the top inhibitors indicating the important roles in acquiring the hydrogen bonds to these residues in the main protease’s binding process.

In the top 10 inhibitors as listed in Table 1, T9J in the complex 5rg1 is only one non-covalent inhibitor. The rest belongs to the class of targeted covalent inhibitors (TCI) in which they interacts with the protein residues, i.e., cysteine, to form a covalent complex strongly neutralizing target’s function. However, the major disadvantage of TCIs is the association with the high toxicity risks. TCIs’ strong covalent bond can irreversibly modify the unintended protein targets in the human body. As a result, the top covalent inhibitors in SARS-CoV PDB-noBA dataset may have little chance to become approved
market drugs in comparison to their non-covalent counterparts such as T9J in 5rg1.

Due to the popularity of the binding site P1 among 137 interested inhibitors, we mainly analyze the interaction network around the residues in that region. Out of 110 molecules binding to P1, there are 103 inhibitors forming at least one hydrogen bond to the nearby amino acid in the SARS-CoV-2 main protease. We have identified 20 different residues in the binding pocket P1 composing hydrogen bonds to these small molecules. Fig. 4 illustrates the frequency of these 20 residues across 110 inhibitors. Based on Fig. 4, Gly143 residue is the most attractive site to form the hydrogen bond. It appears in 53.6% of 110 intermolecular bonding interactions, followed by Glu166 residue with a frequency of 39.1%; residue Cys145 and His163 also occupy 38.2% and 30.9%, respectively. It is worth noting when these molecules form a hydrogen bond with Cys145, they also constitute another hydrogen bond with Gly143. In all cases, both these residues share the same

Table 2  Interaction analysis in the binding pockets of top 4 complexes in term of binding affinity predicted by our MathDL models

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>Ligand ID</th>
<th>Hydrogen bond</th>
<th>Covalent bond</th>
</tr>
</thead>
<tbody>
<tr>
<td>7c8t</td>
<td>Nol</td>
<td>His163, His164, Cys145, Gln189, Gly143, Glu166</td>
<td>Cys145</td>
</tr>
<tr>
<td>5rg1</td>
<td>T9J</td>
<td>His163, Glu166</td>
<td>Cys145</td>
</tr>
<tr>
<td>6xhm</td>
<td>V2m</td>
<td>His163, His164, Cys145, Gln189, Phe140</td>
<td>Cys145</td>
</tr>
<tr>
<td>7bay</td>
<td>N3</td>
<td>His163, His164, Cys145, Gln189, Thr190, Glu166, Phe140, Gly143</td>
<td>Cys145</td>
</tr>
</tbody>
</table>

Fig. 3  The interactions between the top 4 inhibitors in the SARS-CoV PBD-noBA dataset and SARS-CoV-2 M\textsuperscript{pro}: (a) 7c8t; (b) 5rg1; (c) 6xhm; and (d) 7bay. Inhibitors are shown in the purple color. Hydrogen bonds are marked in dashed green lines, and covalent bonds are depicted in solid blue lines. All interactions are shown with the distance information in Å.
hydrogen-bond acceptor. Besides the hydrogen bond network, 71 ligands in the SARS-CoV PDB-noBA dataset form a covalent bond to $\gamma$-sulfur of Cys145. Except the second one, all the others in the top 10 inhibitors are equipped with that covalent bond (see Table S8 in ESI†).

Furthermore, we are interested in the binding energy distribution associated with the interaction network. Fig. 5 depicts the violin plot of that distribution across four categories, namely no H-bond (no hydrogen bond), H-bond (at least one hydrogen bond), no cov. bond (no covalent bond), and cov. bond (at least one covalent bond). Hydrogen bond interactions that are expected to play an important role in the binding mechanism are well captured in our MathDL models. Specifically, while the average energy of inhibitors having no hydrogen bond is $-6.62 \text{ kcal mol}^{-1}$, the average energy of ones with hydrogen bond is as low as $-7.23 \text{ kcal mol}^{-1}$.

It is noted that our MathDLs only measure the non-covalent binding affinity. The covalent bond appearing at the final covalent complex is not properly accounted for in our framework. Therefore, it is expected that our models sometimes overestimate the covalent-bond inhibitors over the non-covalent-bond candidates. Fig. 5 reveals molecules in the group of covalent bonds generally are predicted with lower binding energy with an average being $-7.42 \text{ kcal mol}^{-1}$ in comparison to $-6.89 \text{ kcal mol}^{-1}$ averagely measured on ones without covalent bonds.

### 2.2.3 Fragment analysis

To design the lead molecules, it is of importance to have promising fragments from existing inhibitors against the drug targets. Therefore, in the present work, we study all the fragments decomposed from 110 inhibitors attached to the binding site $P_1$. To carry out this task, we utilize BRICS algorithm via RDkit. In BRICS model, there are 16 chemical environments indicated by linkers denoted by $L_1$, $L_2$, ..., $L_{16}$. The BRICS decomposition gives rise to a total of 185 unique fragments, which are all presented in Table S9 in ESI†. Fig. 6 illustrates top 12 common fragments in terms of their frequencies. Noting that the second frequent fragment, $L_1$-C(C)==O, often constitutes a hydrogen bond with Gly143 and in many cases forms a covalent bond with Cys145.

### 3 Materials and methods

#### 3.1 Datasets

Our deep learning-based scoring function, MathDL, was trained on public databases including PDBbind and ChEMBL. The PDBbind sets contain all complexes with crystal structures deposited in the PDB with the binding affinities not limited to $K_d$, $K_i$, and IC$_{50}$ reported in the literature. In this work, we employ the PDBbind v2019, the latest version of its generation. The v2019 version of the PDBbind consists of 17 679 protein–ligand complexes. However, the data preprocessing of the MathDL only retains 17 382 complexes. Among them, there are 10 485 ligands measured on $K_i$ and 6537 ligands measured in IC$_{50}$.

ChEMBL is another manually curated database of bioactive molecules. Currently, ChEMBL contains more than 2 million compounds in the SMILES string format. Excluding 30 main protease inhibitors in PDBbind data, we have found other 277 small molecules on ChEMBL with reported $K_d$/IC$_{50}$. Additionally, we have found more than 300 other SARS-CoV main protease inhibitors from literatures. In total, there are more than 600 ligands bound to SARS-CoV/SARS-CoV-2 main protease having the experimental binding affinities; among them, there are 44 crystal structures. For compounds without the crystal structures, MathPose is utilized to generate their 3D conformations. The predicted 3D coordinates of these structures are presented in the SDF format and available in ESI†. Currently, there are roughly 137 ligands forming crystal complexes with the SARS-CoV-2 main protease on PDB without the report of the experimental inhibitor activities. Most of them are deposited by the PanDDA analysis group (https://pandda.bitbucket.io/#/).

To serve model validation purposes, we classify the selected data into five different groups as listed in Table 3. Specifically, PDBbind v2019 is the biggest set in this compilation with its PDB IDs and experimental binding affinities listed in Table S1 in ESI†. PDBbind v2016 core set is a subset of PDBbind v2019 and is formed by 290 complexes representing all protein classes in the refined set of PDBbind v2016†. The PDB IDs of all
complexes in the PDBbind v2016 core set are provided in Table S2.† We also collect all Mpro complexes of SARS-CoV/SARS-CoV-2 on the PDB, denoted by SARS-CoV PDB, which results in a total of 192 structures (see Table S3†). Among them, there are 44 ligands with the report of experimental binding affinities denoted by SARS-CoV PDB-BA (see Table S4†). Furthermore, we are interested in the set of SARS-CoV-2 Mpro complexes in the aforementioned SARS-CoV PDB set but their affinities are not presented or undisclosed. We call this set SARS-CoV PDB-noBA with PDB IDs listed in Table S5.† To enrich our training data targeting SARS-CoV/SARS-CoV-2 main protease inhibitors, we gather some inhibitors reported on the literature.4,25 For those compounds with only 2D information, we limit ourselves to ones having the similarity score based on the path-based fingerprint FP2 no lower than 0.6 to at least one inhibitor in the SARS-CoV PDB set. As a result, we arrive at a set of 141 structures named SARS-CoV 2D (see Table S6†). Combining SARS-CoV PDB-BA and SARS-CoV 2D data sets, we finalize a reliable database focusing on SARS-CoV/SARS-CoV-2 main protease inhibitors. Notice that the binding affinities in this set are all reported in IC50. Table S7 in ESI† presents the PDB IDs as well as the experimental binding energies of these ligands.

### 3.2 Methods

#### 3.2.1 MathDL

The MathDL models developed in this work are reformulated from our early model bearing the same name. MathDL was designed for the prediction of various druggable properties of 3D molecules.4 In the past three years, MathDL has been proved to be the top competitor in D3R Grand Challenges.
**Algebraic topology-based representations.** Algebraic topology studies the topological spaces with the use of abstract algebra, which can dramatically simplify the geometric topology studies. Persistent homology (PH) is one of the algebraic topologies, which can dramatically simplify the geometric topology studies. Persistent homology (PH) is one of the algebraic topologies, which can dramatically simplify the geometric topology studies.

**Chain complex.** A formal sum of q-simplices in simplicial complex $K$ with coefficients in an algebraic field (typically $\mathbb{Z}_2$) is a $q$-chain. A set of all $q$-chains of the simplicial complex $K$ equipped with an algebraic field is called a chain group and denoted as $C_q(K)$. The boundary operator is defined by $\partial_q : C_q(K) \to C_{q-1}(K)$ to relate the chain groups. More specifically, we denote $\sigma_q = [v_0, v_1, ..., v_q]$ for the $q$-simplex spanned by its vertices, and then the boundary operator can be represented as:

$$\partial_q \sigma_q = \sum_{i=0}^{q} (-1)^i \sigma_{q-1}^i.$$  

Here, $\sigma_{q-1}^i = [v_0, ..., v_{i-1}, v_{i+1}, ..., v_q]$ is the $(q-1)$-simplex with $v_i$ being omitted. The sequence of chain groups connected by boundary operators is called the chain complex and expressed as:

$$\cdots \xrightarrow{\partial_{q+2}} C_{q+1}(K) \xrightarrow{\partial_{q+1}} C_q(K) \xrightarrow{\partial_q} C_{q-1}(K) \xrightarrow{\partial_{q-1}} \cdots$$

The $q$-cycle group $Z_q(K)$ and the $q$-boundary group $B_q(K)$ are defined as $Z_q(K) = \ker(\partial_q) = \{ c \in C_q(K) | \partial_q c = 0 \}$ and $B_q(K) = \text{im}(\partial_{q+1}) = \{ \partial_{q+1} c | c \in C_{q+1}(K) \}$. The $q$-th homology group is the quotient group $H_q(K) = Z_q(K) / B_q(K)$. Moreover, the rank of $q$-th homology group can be computed as $\text{rank}H_q(K) = \text{rank}Z_q(K) - \text{rank}B_q(K)$, which is denoted as the $q$-th Betti number $\beta_q$. To be noticed, the $q$-th Betti number count the number of $q$-dimensional holes that can not be continuously deformed to each other.

**Persistent homology.** A filtration of a simplicial complex $K$ is a nested sequence of subcomplexes of $K$ such that $\emptyset = K_0 \subseteq K_1 \subseteq K_2 \subseteq ... \subseteq K_m = K$. Then the $p$-persistent $q$th homology group of $K_i$ is defined as:

$$H_q^p(K_i) = Z_q(K_i) / (B_q(K_{i+p}) \cap Z_q(K_i)).$$  

**Fig. 7** A framework of MathDL energy prediction model which integrates advanced mathematical representations with sophisticated CNN architectures.
Here the rank of $H_p^g(K_i)$ counts the number of $g$-dimensional holes in $K_i$ that are still alive in $K_i^{rep}$, which is called the $p$-persistent $g$th Betti number. The persistent homology not only records the topological information at a specific configuration, but also tracks the changes along with the filtration parameters. More specifically, the topological changes will be preserved in the persistent barcodes. In MathDL, we make use of the persistent homology barcodes by dividing them into bins and calculating the birth, death, and persistence incidents in each bin to enrich our algebraic topological representations.

### 3.2.1.2 Element specific considerations.

The protein–ligand complex is structural and also biological. The persistent homology provides a theoretical approach to encode high-dimensional spatial data of protein–ligand complexes into algebraic topological representations. In this section, we address the biological considerations for biomolecular complexes. There are many kinds of interactions that exist in the protein–ligand complex, such as electrostatics, hydrogen bonds, and hydrophobic effects. Although persistent homology can capture the interactions among the nearest neighbors, the long-range interactions will be hindered. This difficulty can be avoided via the deployment of the element-specific attention. Here the rank of $K_i$ considerations.

There are 4 commonly atom types in protein, namely C, N, O, S, and there are 11 commonly atom types in ligand, including C, N, O, S, P, F, Cl, Br, I, H, B. We include Boron in the ligand atom type consideration since it appears in more than 200 small molecules. In each fold prediction task, MathDL trains on the partial data of SARS-CoV BA in conjunction with overlapped folds. In each fold prediction task, MathDL trains on the most optimal model for this benchmark, MathDL is trained for 1000 epochs. Then, we pick the epoch based on the root-mean-squared error (RMSE) of the PDBbind v2016 core set prediction. We have found that MathDL achieves the smallest RMSE in this experiment at 140 epochs. Specifically RMSE, $R_p$, and $\tau$ metrics on the v2016 core set are 1.56 kcal mol$^{-1}$, 0.858, and 0.671, respectively. Meanwhile, the training accuracy is 0.387 kcal mol$^{-1}$ in terms of RMSE and its Pearson’s correlation coefficient is $R_p = 0.994$. These performances reveal that our MathDL converges very fast and with only 140 epochs and maintains a good balance between training and testing accuracies. This is a state-of-the-art performance since our MathDL is ranked in the second place in comparison to 33 other scoring functions (see Fig. 8). It is noted that the top model is TopBP con, published in our previous work with $R_p = 0.861$. TopBP con is the consensus of gradient boosted tree and deep learning-based models. If only the deep learning framework is considered, the performance of TopBP (denoted by TopBP-DL) on the core set of PDBbind v2016 is $R_p = 0.848$.

It is worth mentioning that except for our MathDL, all machine learning-based scoring functions listed in Fig. 8 were trained on the PDBbind v2016 refined set of 3767 complexes. As mentioned above, the current MathDL is compiled on a much larger training set comprised of 17,211 complexes selected from PDBbind v2019 and SARS-CoV BA data. Even the current MathDL has not outperformed its predecessor, i.e., TopBP con, MathDL is still a preference model since it is trained on a diverse data set covering various protein families and different binding energy ranges. As a result, it is expected to deliver more reliable predictions on the SARS-CoV-2 inhibitor, especially when this main protease family is not included in the training data of previous TopDL models. The resulting MathDL model is labeled as MathDL-Core2016 and is utilized to predict affinities of complexes in SARS-CoV PDB-noBA in Section 2.1.

#### 3.3 Validation

##### 3.3.1 PDBbind v2016 core set benchmark.

In this validation task, we will testify our model against 290 complexes in the PDBbind v2016 core set. This is a prevalent test set to assert the scoring ability of a binding affinity prediction model and has attracted lots of research groups to devote the effort to improve the Pearson’s correlation coefficient ($R_p$) and Kendall’s tau ($\tau$) on this core set performance.\cite{16,17,18,19} In the current work, we merge the PDBbind v2019, SARS-CoV PDB-BA, and SARS-CoV 2D sets but removing the duplicates and excluding the PDBbind v2016 core set complexes to attain a training set of 17,211 complexes. MathDL with the architecture described in Section 3.2.1 is trained on those complexes. The resulting model is utilized to predict the binding affinity of 290 structures in the PDBbind v2016 core set.

With the purpose of exploring the most optimal model for this benchmark, MathDL is trained for 1000 epochs. Then, we pick the epoch based on the root-mean-squared error (RMSE) of the PDBbind v2016 core set prediction. We have found that MathDL achieves the smallest RMSE in this experiment at 140 epochs. Specifically RMSE, $R_p$, and $\tau$ metrics on the v2016 core set are 1.56 kcal mol$^{-1}$, 0.858, and 0.671, respectively. Meanwhile, the training accuracy is 0.387 kcal mol$^{-1}$ in terms of RMSE and its Pearson’s correlation coefficient is $R_p = 0.994$. These performances reveal that our MathDL converges very fast and with only 140 epochs and maintains a good balance between training and testing accuracies. This is a state-of-the-art performance since our MathDL is ranked in the second place in comparison to 33 other scoring functions (see Fig. 8). It is noted that the top model is TopBP con, published in our previous work with $R_p = 0.861$. TopBP con is the consensus of gradient boosted tree and deep learning-based models. If only the deep learning framework is considered, the performance of TopBP (denoted by TopBP-DL) on the core set of PDBbind v2016 is $R_p = 0.848$.

It is worth mentioning that except for our MathDL, all machine learning-based scoring functions listed in Fig. 8 were trained on the PDBbind v2016 refined set of 3767 complexes. As mentioned above, the current MathDL is compiled on a much larger training set comprised of 17,211 complexes selected from PDBbind v2019 and SARS-CoV BA data. Even the current MathDL has not outperformed its predecessor, i.e., TopBP con, MathDL is still a preference model since it is trained on a diverse data set covering various protein families and different binding energy ranges. As a result, it is expected to deliver more reliable predictions on the SARS-CoV-2 inhibitor, especially when this main protease family is not included in the training data of previous TopDL models. The resulting MathDL model is labeled as MathDL-Core2016 and is utilized to predict affinities of complexes in SARS-CoV PDB-noBA in Section 2.1.

##### 3.3.2 5 fold cross-validation on SARS-CoV BA set.

In this section, we test the performance of our MathDL against 185 inhibitors in the SARS-CoV BA set aforementioned in Table 3. Among those ligands, there are 44 X-ray crystal structures and the rest are in 2D SMILES strings. We employ MathPose to predict 3D structures of those 2D ligands. To carry out the validation, we randomly split the SARS-CoV BA set into 5 non-overlapped folds. In each fold prediction task, MathDL trains on the partial data of SARS-CoV BA in conjunction with PDBbind v2019 set. This situation results in two different ways
of training our MathDL model. The first approach is a traditional MathDL architecture with the training set combining both SARS-CoV BA and PDBbind v2019 complexes. The second model makes use of multi-task learning. In each epoch, the weights of the MathDL architecture are learned through the information from PDBbind v2019 set, then only the fully connected layers are trainable when learning SARS-CoV BA structures. Finally, we come up with 10 different MathDL models in which the traditional MathDL frameworks are labeled as MathDL-All-i and multi-task MathDL is named MathDL-MT-i with i running from 1 to 5. In each model, after 100 epochs, we start monitoring which epoch that helps our model achieve the smallest RMSE on the test set. Table 4 reveals that MathDL-All models are well trained with the averaged accuracy RMSE = 0.286 kcal mol\(^{-1}\), Pearson’s correlation coefficient \(R_p = 0.994\), and Kendall’s tau \(\tau = 0.934\). Their averaged performances on test data across 5-fold of the SARS-CoV BA set are found to be \(R_p = 0.729, \tau = 0.540,\) and RMSE = 0.789 kcal mol\(^{-1}\). These results endorse the reliability of these models in the binding affinity prediction of SARS-CoV/SARS-CoV-2 inhibitors. Table 4 also lists the training and testing performances of five multi-task learning models. The averaged training performance of the MathDL-MT model is \(R_p = 0.995, \tau = 0.941\) and RMSE = 0.275 kcal mol\(^{-1}\). The accuracy of the multi-task architecture on the test sets is similar to MathDL-All with \(R_p = 0.727, \tau = 0.532,\) and RMSE = 0.822 kcal mol\(^{-1}\). With these promising results, it is encouraging to carry out MathDL models to predict unknown binding affinities of SARS-CoV/SARS-CoV-2 inhibitors. It is worth noting that if the 5-fold cross-validation is conducted purely on the SARS-CoV BA set, the average \(R_p, \tau\) and RMSE are as low as 0.561 and 0.388, respectively. These results strongly support the inclusion of diverse information such as PDBbind v2019 in conjunction with sophisticated deep learning architectures to achieve the accurate binding energy prediction of M\(^{\text{pro}}\) inhibitors.

### 4 Conclusion

SARS-CoV-2 main protease (M\(^{\text{pro}}\)) is the most favorable target for COVID-19 drug discovery due to its conservative nature and low similarity with human genes. Structure and binding affinity of protein-drug complexes are of paramount importance for understanding the molecular mechanism in drug discovery. However, there are only two SARS-CoV-2 M\(^{\text{pro}}\) inhibitor structures available with binding affinities, highlighting current challenges in COVID-19 drug discovery.

This work presents the reliable binding affinity prediction and ranking of 137 M\(^{\text{pro}}\)-inhibitor crystal structures that have no reported experimental binding affinity. We first curate a set of more than 600 M\(^{\text{pro}}\) inhibitors with binding affinities from public resources, such as PDBbind, ChEMBL and the scattered literature. Among these inhibitors, 141 are retained based on their high similarity with available M\(^{\text{pro}}\)-inhibitor complex structures and built with three dimensional (3D) poses using
our MathPose. Together with 44 another SARS-CoV or SARS-CoV-2 Mpro–inhibitor complexes, we compose a training set of 185 reliable SARS-CoV-2 Mpro–inhibitor complexes. Our earlier MathDL models are reformulated with algebraic topology to accommodate 119 new complexes and 17 382 complexes from the PDBbind v2019 general set in both single-task and multitask settings, which have never been available before. The resulting MathDL models are rigorously validated via PDBbind v2016 core set benchmark in which it outperforms state-of-the-art models in the literature. Most importantly, our MathDL achieves promising cross-validation accuracies on the SARS-CoV family inhibitors with the averaged Pearson’s correlation coefficient as high as 0.73.

Additionally, the present work unveils that Gly143 of Mpro is the most attractive region to form hydrogen bonds, followed by Glu166, Cys145, and His163. There are 71 inhibitors interacting with SARS-CoV-2 Mpro to form covalent complexes. Those covalent bonds are mostly composed between dicarbon monoxide groups in inhibitors and γ-sulfur on Cys145. There are only one non-covalent complex in our top 10 ranked, namely 5rg1. To provide a potential resource for lead molecule design, we employ the BRICS algorithm to decompose all the inhibitors of the prominent binding site on Mpro and obtain 185 unique fragments.

The predicted binding affinities and their ranking of 137 Mpro–inhibitor crystal structures, the bonding analysis, and the fragment decomposition have significantly extended current knowledge and understanding of SARS-CoV-2 Mpro and inhibitor interactions and, thus offered valuable information toward COVID-19 drug discovery.

Conflicts of interest
There are no conflicts to declare.

Acknowledgements
This work was supported in part by NIH grant GM126189, NSF Grants DMS-1721024, DMS-1761320, and IIS1900473, Michigan Economic Development Corporation, George Mason University award PD45722, Bristol-Myers Squibb, and Pfizer. The authors thank The IBM TJ Watson Research Center, The COVID-19 High Performance Computing Consortium, and NVIDIA for computational assistance.

References


