

# Transferable Graph Neural Fingerprint Models for Quick Response to Future Bio-Threats

Wei Chen<sup>†§</sup>, Yihui Ren<sup>\*§</sup>, Ai Kagawa<sup>\*</sup>, Matthew R. Carbone<sup>\*</sup>, Samuel Yen-Chi Chen<sup>\*</sup>, Xiaohui Qu<sup>†</sup>, Shinjae Yoo<sup>\*</sup>, Austin Clyde<sup>‡</sup>, Arvind Ramanathan<sup>‡</sup>, Rick L. Stevens<sup>‡</sup>, Hubertus J. J. van Dam<sup>¶</sup>, and Deyu Lu<sup>†</sup>

<sup>†</sup> Center for Functional Nanomaterials, Brookhaven National Laboratory, Upton, NY, USA, dlu@bnl.gov

<sup>\*</sup> Computational Science Initiative, Brookhaven National Laboratory, Upton, NY, USA, yren@bnl.gov

<sup>‡</sup> Data Science and Learning Division, Argonne National Laboratory, Lemont, IL, USA

<sup>¶</sup> Condensed Matter Physics & Materials Science, Brookhaven National Laboratory, Upton, NY, USA, hvandam@bnl.gov

<sup>§</sup> Authors contributed equally.

**Abstract**—Fast screening of drug molecules based on the ligand binding affinity is an important step in the drug discovery pipeline. Graph neural fingerprint is a promising method for developing molecular docking surrogates with high throughput and great fidelity. In this study, we built a COVID-19 drug docking dataset of about 300,000 drug candidates on 23 coronavirus protein targets. With this dataset, we trained graph neural fingerprint docking models for high-throughput virtual COVID-19 drug screening. The graph neural fingerprint models yield high prediction accuracy on docking scores with the mean squared error lower than 0.21 kcal/mol for most of the docking targets, showing significant improvement over conventional circular fingerprint methods. To make the neural fingerprints transferable for unknown targets, we also propose a transferable graph neural fingerprint method trained on multiple targets. With comparable accuracy to target-specific graph neural fingerprint models, the training and data efficiency of the transferable model is several times higher. We highlight that the impact of this study extends beyond COVID-19 dataset, as our approach for fast virtual ligand screening can be easily adapted and integrated into a general machine learning-accelerated pipeline to battle future bio-threats.

**Index Terms**—Graph Neural Networks, Transfer Learning, Bioinformatics

## I. INTRODUCTION

The knowledge of the protein-ligand interaction is essential to many fields in the life sciences, such as biophysics, structural bioinformatics and drug discovery [1]. Detailed information on the atomic structures and energetics of the protein-ligand complex in the docking conformation is key to unravelling the docking mechanism, which is governed by multiple factors including, e.g., shape matching, electrostatics, hydrogen bonding and van der Waals forces.

To understand the specific action of a protein on a substrate the lock-and-key model was first proposed by Fischer in 1894 [2]. The lock-and-key model proved useful in a variety of contexts including protein-protein interactions as well as protein-ligand interactions. With the emergence of increasingly capable high performance computers, it became possible to automate the search for optimal protein-ligand alignments. This led to the first docking code being developed in 1982 [3]. In addition, it was realized that the lock-and-key model could be

used for rational drug design [4]. This realization quickly led to the design and deployment of docking programs specifically for this purpose [5]. Despite remaining challenges, thanks to the growing computing capabilities and improvements in methods and software, docking became a key component in rational drug design.

One outstanding challenge is that the accuracy of the docking results is limited by the approximations required to accelerate the method. These approximations involve the way the scoring function accounts for entropy and desolvation effects [6]. Other limitations stem from the extent to which docking allows for the flexibility of the protein [6]. These approximations are to a degree due to the scale of the problem that needs to be solved. While docking a single ligand using empirical force field can be done in seconds, the chemical space of drug-like molecules is vast. The size of this space for small molecules with up to 30 atoms has been estimated to be of the order of  $10^{60}$  molecules [7]. While this formidably large chemical space can never be fully explored, smaller but still huge subsets have recently explicitly been considered. The GDB-17 database has 166 billion molecules [8]; ZINC15 contains over 750 million purchasable compounds [9]; ENAMINE enumerates over 22.7 billion compounds, with a database of 4.5 billion *REAL* molecules for download [10]. Exploring such large molecular spaces is still very expensive, even using fast docking programs, calling for yet more efficient screening techniques.

One promising path forward is to develop machine learning (ML) models that predict the docking score directly from ligand structures or ligand-target protein structure complexes [11], [12]. Early efforts along these lines have been made since at least 2010 using multiple linear regression, partial least squares regression, random forests, support vector machines, and artificial neural networks [13]. Since then, a variety of machine learning models have been developed, including multiple linear regression [14], multivariate adaptive regression splines [14], gradient boosted trees [15], [16], boosted regression trees [14], random forests [17], [18], k-nearest neighbors [14], [19], support vector machines [14], [19], logistic regression [18], [19], artificial neural networks [15],

[20], convolutional neural networks [21]–[23], and graph convolutional neural networks [24], [25].

Other than the details of the machine learning methods, there are several distinct differences in these models, such as the choice of ground truth in the training set and target applications. For models aimed at generic binding properties, common choices for ground truth on experimental data such as PDBBind [26] are used [13], [14], [18], [21]; Binding MOAD [27] and Astex [28] have also been considered [29]. While highly valuable, these experimentally obtained datasets tend to be relatively small. For example PDBBind 2020 contains 19,443 experimentally characterized protein-ligand complexes, which is a small number compared to the possible combinations of more than 188,430 experimental structures in the PDB and billions of theoretically available ligands. An alternative way of obtaining larger ground truth datasets is using computational docking programs to generate simulated datasets. For example, Autodock Vina [19], [23], [25], [30] and Gold [17], [31] have been used to provide ground truth data. The caveat with using these datasets is that they are affected by the same limitations that docking scores in general tend to have.

Our work has three contributions: 1) we build a COVID-19 drug docking dataset of about 300,000 drug candidates on 23 coronavirus protein targets; 2) we conduct a systematic study of the popular neural fingerprint models and compare the model performance on this large docking dataset with conventional circular fingerprint models; and 3) we demonstrate the learned neural fingerprints can be used for emerging protein targets under a transfer learning setting.

## II. COVID-19 DOCKING DATASET

Since the first documented case at the end of 2019, coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has quickly evolved into a worldwide pandemic. Unlike previous coronaviruses, such as SARS-CoV and the Middle East Respiratory Syndrome (MERS), SARS-CoV-2 is proved to be significantly more contagious, leading to its exponential spread and a large number of fatalities. With roughly 6.95 million deaths and 768 million infections as of July 2023 [32], the disastrous impact of this pandemic calls for urgent pharmacological progress in, e.g., vaccines, drugs, and interferon therapies to combat COVID-19 [33].

Potential antiviral treatments of SARS-CoV-2 can be divided into two categories acting on either the human immune system or the coronavirus [34]. In this work, we focus on the drug molecule docking study of the latter. In general, virus proteins fall into three categories: 1) structural proteins (SPs), which form the virion particles, 2) non-structural proteins (NSPs), which are involved with the virus replication in the host cell, and 3) accessory proteins, which interfere with the host cell’s innate immune response [35]. A ligand drug molecule may act on SPs to prevent virus from assembling or binding to human cell, or on critical NSPs to inhibit virus RNA synthesis and replication [34].

Since 2020, over 1300 SARS-CoV-2 protein structures (either by themselves or in complexes with other compounds) have been resolved from experimental facilities around the world. The atomic structures of these proteins have been determined to high accuracy, which provide the essential structural information of the drug docking sites on SARS-CoV-2. A detailed explanation of the functions of SARS-CoV-2 proteins in the virus life cycle can be found in a recent review [35]. Among them, proteins of particular interest are PLPro (an NSP3 domain) and 3CLPro (NSP5), which are the proteases that cut the polyproteins encoded by the viral RNA into active proteins, and ADP Ribose phosphatase (also an NSP3 domain), an innate immune response antagonist. The CoV protein is the receptor binding domain that is the part of the spike protein that binds to the ACE2 receptor triggering the infection. NSP9 is an RNA binding protein that is involved in the viral RNA replication although its precise role is still unclear. NSP10 is the co-factor in the NSP16-NSP10 complex that methylates the cap of newly synthesized RNA, an essential step for RNA stability and function. NSP15 is thought to modify viral RNA at 3’ Uracil locations to evade detection by the cell’s innate immune system [36]. In addition, researchers leverage the knowledge of the protein functions in other coronaviruses, such as SARS-CoV and MERS as well as the interaction map of SARS-CoV-2 and human proteins [37]. In this work, we consider 23 pertinent SARS-CoV-2 NSP docking sites as described in Table I.

TABLE I  
DESCRIPTION OF THE 23 SARS-CoV-2 NSP TARGETS.  $n_p$  INDICATES THE NUMBER OF POCKETS.

Protein	$n_p$	Description
3CLPro	1	3C-like protease [38]
ADRP-ADPR	2	ADP-ribose phosphatase in complex with ADP ribose [39]
ADRP	3	ADP-ribose phosphatase [39]
COV	4	Receptor binding domain
NSP9	2	Component of RNA polymerase complex [40]
NSP10	3	Component of 2’-O-RNA methyltransferase complex [41]
NSP15	2	Nidoviral RNA uridylylate-specific endoribonuclease [42]
ORF7A	1	Interferon response antagonist [43]
PLPro monomer	3	Papain-like protease monomer [44]
PLPro dimer	2	Papain-like protease dimer

## III. METHODS

### A. Docking score prediction workflow

The schematics of our workflow is shown in Fig. 1, which includes generating a docking dataset from large scale docking simulations (top, shaded in blue) and training surrogate models (bottom, shaded in purple) that can efficiently screen COVID-19 drug candidates in the vast drug-like molecule space. Our ML model predicts the docking scores of drug candidate molecules. While predicting the docking pose using ML models is an exciting open question, it is beyond the scope of

this study. The technical details of this workflow are explained below.

First, we performed docking simulations of 310,693 compounds, including the drug bank compounds, onto each of the 23 coronavirus protein targets using Autodock [45]. The set of targets was generated by using Fpocket [46] on each protein to identify the top 4 most druggable pockets. The set of pockets was further reduced based on prior experimentally identified binding motifs and visual inspection. Subsequently all the molecules in the ligand set were processed one-by-one by first converting the molecule’s SMILES string to a three dimensional structure using OpenBabel [47] and the subsequent docking simulations using AutodockTools and Autodock 4.2 [48]. Autodock 4.2 uses an empirical force field to estimate the ligand binding free energy, which contains pair-wise terms to calculate the interaction between two molecules and an empirical model to estimate contributions from environmental water [49]. The scoring function in Autodock contains the van der Waals, hydrogen bonding, electrostatic, and desolvation energies [50]. For each ligand-pocket pair the hybrid genetic algorithm and local search (GA-LS) procedure was executed for 20 times. Each procedure went through maximum 2.5 million energy evaluations to find the lowest energy pose. The final 20 poses were clustered at the end of the AutoDock run and ultimately the best one was selected. The simulation outputs include the optimal docking pose and docking score. The datasets of docking scores are available on Github<sup>1</sup>.

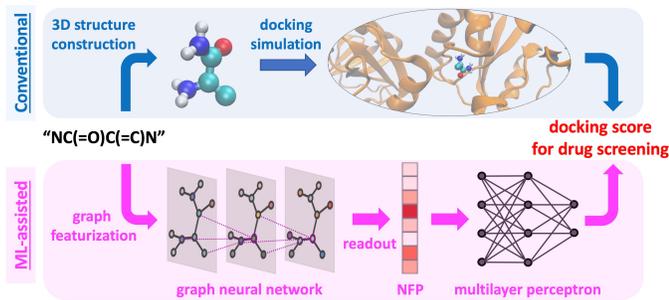


Fig. 1. A schematic flowchart that illustrates two individual processes to compute the docking score of a drug candidate molecule from its SMILES string. The upper route shows the conventional docking simulation, where the docking pose of an exemplary molecule (NC(=O)C(=C)N) on ADRP\_pocket13 is shown. The bottom route shows the graph neural fingerprint model that is trained on the docking simulation data and is able to provide fast and high-fidelity predictions on unknown molecules for drug screening.

## B. Machine learning-based surrogate models

Starting from the SMILES code of molecules, we considered two types of fingerprinting methods based on the featurization of molecules: conventional circular fingerprints (CFP) [51], [52] and neural fingerprints (NFP) [53]. A CFP, such as de Morgan and extended-connectivity fingerprints, abstracts molecular structure information as a vector via hashing, which has been used widely for molecular similarity search

for its fast processing time. It is convenient to use CFPs as the input feature of a neural network to perform molecular property predictions, such as the docking score for a particular docking target as shown in Fig. 2a. Despite the utility for similarity search, the limited structural and chemical information content retained in CFPs impairs their performance for prediction tasks. On the other hand, a NFP-generating model typically has three components: 1) the graph representation and feature embedding, 2) a graph neural network (GNN)-encoder, mapping molecules to a NFP in a fixed-size vector representation, and 3) a multilayer feedforward regressor, e.g., a multilayer perceptron (MLP), to predict a target property. These three components are trained together end-to-end as shown in Fig. 2b. Although it is quite likely that the learned NFP of the same bit-length can outperform the CFP in the docking score prediction task, a systematic comparison is warranted.

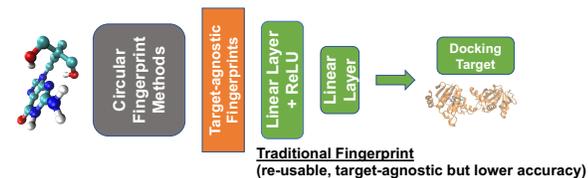
In this study, we conducted a systematic study of five types of molecular fingerprinting methods using both CFP and NFP. For conventional CFP methods, we considered de Morgan circular fingerprint (Morgan) and extended-connectivity fingerprint (ECFP) implemented in the open-source cheminformatics software RDKit [54]. For GNN methods, we considered three popular GNN variations: Gated Graph Convolutional Network (GatedGCN) [55], [56], GraphSAGE [57], [58] and a boilerplate Message-Passing Neural Network (MPNN) [59]. We ensured the fingerprints have the same bit-length of 2048. Specifically, the CFPs have a fixed length of 2048 bits, while the NFPs are represented by a vector of 16-bit floating points of length 128. We found that the NFP model performance remained the same when we reduced the NFP length from 128 to 70 in GatedGCN and GraphSAGE models. Then both types of fingerprints were fed into a 3-layer perceptron for target-specific regression as shown in Figs. 2a and 2b, except for the MPNN where we used only a single layer perceptron.

To process molecular structures into features suitable for GNN training, we first converted the molecular SMILES [60] representation to a multi-attribute undirected graph representation, where nodes and edges correspond to atoms and chemical bonds, respectively, as shown in Figs. 1 and 2b. We followed Gilmer *et al.* [59] to encode the chemical attributes, such as atomic types and bond types, in node and edge features, unless otherwise stated.

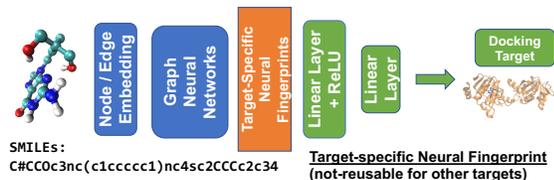
Such categorical features were then mapped to short trainable vectors of real values known as feature embeddings. Therefore, each molecule can be represented as a triplet of matrices of  $(\mathbf{A}, \mathbf{V}, \mathbf{E})$ , where  $\mathbf{A}$  is the adjacency matrix of a graph with added self-loops, and  $\mathbf{V}$  and  $\mathbf{E}$  are node and edge embedding matrices, respectively.

We explored various GNNs [53], [61] to encode a graph topology  $\mathbf{A}$ , its node embedding  $\mathbf{V}$  and edge embedding  $\mathbf{E}$  into a fixed-length fingerprint. GNNs are among the best deep learning models for handling networked data due to its permutation invariance property. Namely, the order of nodes represented in an adjacent matrix does not affect the prediction. Most GNN models consist of an iterative sequence of

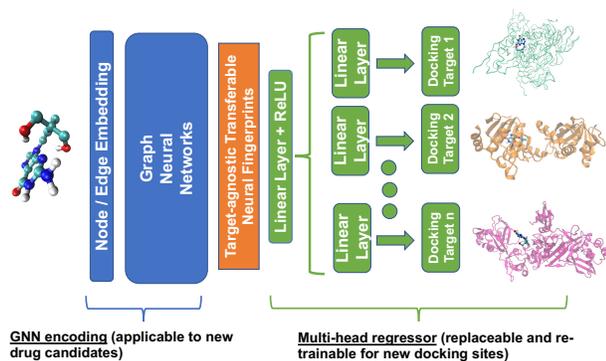
<sup>1</sup>See our dataset on [https://github.com/BC3D/BC3D\\_2021](https://github.com/BC3D/BC3D_2021)



(a) Traditional Circular Fingerprint.



(b) Target-specific Neural Fingerprint.



(c) Transferable Neural Fingerprint.

Fig. 2. Comparison of three fingerprinting schemes: traditional circular fingerprint, target-specific neural fingerprint and transferable neural fingerprint.

alternating two steps: *communication* and *aggregation*. During communication, each node will gather its neighboring node embeddings and incident edge embeddings of previous step. During aggregation, trainable functions (e.g. neural networks) are applied to these embeddings, aggregated into a single vector and used to update the node’s existing embedding. The same procedure applies to the edge embedding updates. Different types of GNNs differ in the functions and procedures applied to the neighboring embeddings and the type of aggregation used.

- Our MPNN model is a slight modification of that presented in Gilmer *et al.* [59]. It is implemented using Deep Graph Library [62], where atoms are featurized according to the Weave atom featurizer [63]. Messages are learned by a MLP, which are passed between neighboring atoms and used to update both the node and edge embeddings using a Gated Recurrent Unit.
- GraphSAGE [55], [56] is a stochastic generalization of graph convolutions which features sampling and hierarchical aggregation. The aggregated embedding is concatenated with that of the central node before a fully-connected layer. We used a max pooling approach as the aggregation function.

- In the Gated-GCN network [55], [56], before aggregation each neighboring node embedding is gated (softmax) by a trainable linear combination of both nodes.

### C. Transferable neural fingerprint

CFP and NFP have an important distinction in the nature of their fingerprints (see Figs. 2a and 2b). The molecular fingerprint in CFP is derived solely from the molecular structure, making it target agnostic and reusable. In contrast, the fingerprint used in the single target NFP model (Fig. 2b) is target specific, as they are trained as part of the neural network with the knowledge of the docking scores of a given target. As a result, the task-specific NFP approach requires re-training for different docking targets. Therefore, *standard NFP methods lack the most prominent advantage of the CFP: pre-computable and target-agnostic*. This drawback severely limits NFP’s practical utility, as for each new protein target it has to be retrained on large amount of docking data and the fingerprint database grows with the number of protein targets, comparing to the case of CFP where the database is invariant to new protein targets. On the physical ground, the docking simulation searches for the lowest energy configuration of the ligand under the given potential field (e.g., electrostatics, van der Waals and hydrogen bond) created by the binding pocket. Therefore, in principle a molecular fingerprint is transferable, as far as it is featurized to encode essential chemical attributes (e.g., atomic charge, van der Waal radius and interaction strength, and the location of hydrogen-bond donor or acceptor) of the ligand that determine the ligand-pocket interaction energy under the target’s potential field.

To this end, we propose transferable neural fingerprints (TNFPs) that combine the benefits of both CFPs and NFPs. On one hand, TNFPs are re-computable and target-agnostic like CFPs; on the other hand, they encode more complete molecular structural information like NFPs. As shown in Fig. 2c, we trained TNFPs via a multi-target model. The learned TNFP can be stored in a database, and the GNN encoder can extract a TNFP from newly added drug molecules. For the newly identified docking target, we can train a dedicated MLP regressor starting from TNFP, which is a much faster and more data-efficient (i.e., requiring less training data) process. To make a fair comparison, all the fingerprints in this study have the same bit length of 2048, either 2048 bits as in CFP or 128 float-16 in NFPs and TNFPs.

## IV. ANALYSIS OF THE DRUG-LIKE LIGAND DOCKING SCORE DATASET

Our raw dataset contains molecular docking results (atomic coordinates for each ligand-target pair from its lowest-energy docked conformation and the corresponding docking score) of 310,693 ligands on each of 23 targets. First we conducted a data cleaning process to retain data relevant to drug screening. Out of the 23 targets, 5 of them (i.e., NSP10\_pocket1, NSP10\_pocket3, NSP10\_pocket26, NSP15\_pocket2, and PL-Pro\_chainA\_pocket4) show positive docking scores on nearly all ligands (more than 99.8%). These target were removed

from the dataset, due to the overall low drug affinity. We also removed about 1K non-drug ligands that are either too large (e.g., protein and RNA) or too small (e.g., salt), as well as nearly 5K non-bonded ligands that contain disconnected parts. After the cleaning process the final dataset includes docking scores of 300,457 ligands on each of 18 different druggable targets.

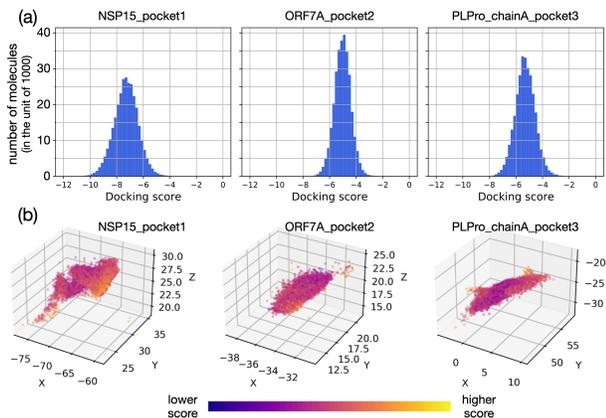


Fig. 3. (a) Docking score distributions and (b) center-of-mass coordinate distributions of molecules for three representative docking targets. The docking scores are in units of kcal/mol and the coordinates in angstrom.

Next we examined the data variance from two perspectives. First, on each target the ligand molecules exhibit a broad range of docking scores (Fig. 3a and the supporting information Fig. S1). The docking score distribution is typically Gaussian-like with a single peak, except for one target – 3CLPro\_pocket1, which shows a bimodal distribution. Second, across different targets the peak position and width of the distribution vary quite significantly. The latter can be also seen from the variation of the average score (i.e., averaged over either all molecules or top 100 molecules in the lower end of the docking score) across the 18 targets (see the supporting information Fig. S2). Since a more negative docking score means stronger ligand-target affinity, targets with more negative tails, including ADRP-ADPR\_pocket5, ADRP-ADPR\_pocket1, and PLPro\_pocket50, are likely promising druggable sites.

To understand the bimodal distribution in 3CLPro\_pocket1, we further examined the docking configurations of the ligand-target pairs. Given more than 300K molecules in the docking simulations, we computed the center-of-mass (COM) distribution of molecules as a proxy of their docking configurations (see Fig. 3b and the supporting information Fig. S3). 3CLPro\_pocket1 is clearly an exception, because its ligand COM distribution shows disconnected regions in space. We sampled and examined structures from each region and found that the pre-defined docking area in 3CLPro\_pocket1 is highly solvent-exposed and consisted of multiple dockable regions. Such observations indicated that the docking area in 3CLPro\_pocket1 may be ill-defined. In contrast, the spatial distributions in all other targets are continuous despite of irregular shape in several targets.

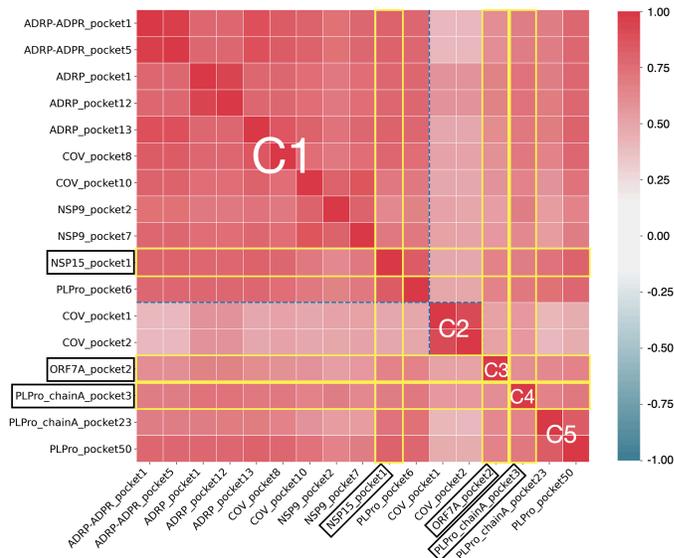


Fig. 4. Hierarchical clustering of targets based on the correlation matrix that measures the similarities between targets. The color indicates the value of the Pearson’s correlation coefficient. Five clusters are labeled as C1-C5, indicated by blocks separated by dashed lines. Targets within the same block belong to the same cluster. Boxed target names indicate the targets chosen to test our multi-target model.

To measure the similarities among targets, we first defined a 300457-dimension vector for each target that contains docking scores of all the ligand molecules. Then we calculated the Pearson’s correlation coefficients and performed a hierarchical clustering analysis on the correlation matrix using the scikit-learn package [64] (see Fig. 4). The targets are all positively correlated, namely, a molecule binds strongly (weakly) to one target is also more likely to bind to other targets strongly (weakly). Based on the correlation matrix the targets were grouped into five clusters (C1 to C5). The largest cluster in the upper left of Fig. 4 contains 11 targets, while the rest of the clusters have at most 2 targets. Within each cluster the targets share significant similarities.

## V. RESULTS AND DISCUSSIONS

We trained two types of target-specific GNN models to predict docking scores of drug candidate molecules – one based on CFP (Fig. 2a) and the other based on NFP of molecules (Fig. 2b). The performance of the prediction on the test set was reported in Table II. As a reference for comparison, the mean squared error (MSE) of the baseline model was also reported, which refers to the error with respect to the average score in the training set. More information about the architectures and configurations of our graph neural networks is explained in the supporting information.

First, all GNN models show significant improvement over the baseline by 2- to 10-fold, suggesting that GNN-based surrogate models can indeed capture the non-trivial correlation between the molecular structure of the ligand and its docking score on a specific target. Second, NFP-based models outperform the CFP-based models systematically with smaller MSE

values by 0.01  $\sim$  0.08. Third, the three NFP-based models perform equally well with minor MSE differences (smaller than 8%) between them, and the GatedGCN model is the best in most cases. We included a scatter plot of the ground truth versus prediction using the GatedGCN model in the supporting information Fig. S4. We noticed that the prediction of docking scores of molecules on 3CLPro\_pocket1 is the worst with the MSE an order of magnitude higher than the rest. This large error is likely caused by the fact that 3CLPro\_pocket1 does not have a well-defined docking pocket as discussed in the above section.

When models are used to perform drug screening, the top ranked (e.g., top 10%) molecules are usually selected for further testing. In this regard the rank correlation between molecules can be more important than the error on an individual molecule when a model is evaluated. A commonly used metric to measure the rank correlation is the concordance index (CI) [65]. In our GatedGCN models, the CIs are between 0.76 and 0.91 (see Fig. S4).

Next we trained a multi-target model via a GNN encoder with multi-head regressors, one for each docking site, as shown in Fig. 2c. We used 14 targets for training and tested the transferability of the learned TNFP on the remaining targets. We excluded 3CLPro\_pocket1 for this task based on the docking score and configuration analysis in the previous section. Based on the similarities among targets (Fig. 4), we chose NSP15\_pocket1, ORF7A\_pocket2, and PL-Pro\_chainA\_pocket3 as the test targets such that they belong to different target clusters. The first target was randomly picked, and the latter two were intentionally chosen to minimize the similarities to the 14 training targets. For the purpose of comparison, we included the results of the single-target GatedGCN and CFP models that were adopted from Table II. On the three test targets, the TNFP model performance shows a noticeable improvement over the CFP models, and is slightly worse than that of GatedGCN (difference within 11.5%), as shown in the supplementary information Fig. S5.

The transferable nature of the TNFP model gives it the advantages in training efficiency and data efficiency. The training efficiency is reflected in the runtime of each model. On average the TNFP model costs 53 seconds/epoch on our GPU node while the GatedGCN single-target model costs almost twice (90 seconds/epoch) and the CFP models about 9 times (410 seconds/epoch) as much as the TNFP model. The training efficiency of TNFP arises from two factors. First, its input dimension, i.e., the number of nodes in the input layer, is much smaller than the CFP models. Second, its graph encoding part is fixed (i.e., transferable) and does not require re-training as opposed to the single-target NFP model.

The data efficiency can be extremely useful when the training data is limited, which is often the case for newly discovered protein targets. As we reduced the training size from the original size (240,000 molecules), the MSE on the test set (size remained 30,457) increases much faster in the GatedGCN model and the two CFP models than the TNFP model (see Fig. 5 for NSP15\_pocket1 and Fig. S6 for the

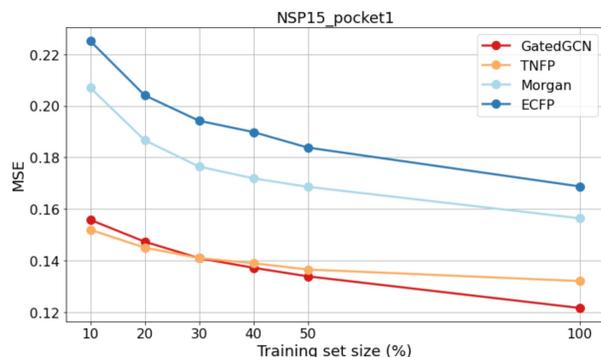


Fig. 5. Training data efficiency of four different models on NSP15\_pocket1.

other two test targets). As we see in Fig. 5, below the crossover at 30%, the TNFP model outperforms the GatedGCN model. With 10% of the training data (24,000 molecules), the MSE of the TNFP model increases by only 15.1% as compared to that of the full training set, while the MSE of the GatedGCN model increases dramatically by 28.1%.

## VI. CONCLUSIONS

In summary, we have built and analyzed a COVID-19 docking datasets consisting of  $\sim 3 \times 10^5$  drug candidates and 23 coronavirus protein targets using Autodock. We have conducted a comprehensive study of various graph neural network methods to construct surrogate models for docking score prediction, including both conventional circular fingerprint methods (ECFP and Morgan) and graph neural fingerprint methods (GatedGCN, GraphSAGE, and MPNN). We found that overall graph neural fingerprint methods outperform the conventional circular fingerprint methods with the same bit-length of 2048, and GatedGCN performs slightly better than GraphSAGE and MPNN. However, graph neural fingerprint methods are target specific and require re-training for new docking targets, which makes them more data intensive and computationally more expensive to train than the conventional circular fingerprint methods. By withholding five representative protein targets as unknown emerging bio-threat, we demonstrated that the neural fingerprints learned via multi-target training exhibits desired target agnostic and reusable properties of circular fingerprints. We found that the transferable graph neural fingerprint model not only outperforms conventional circular fingerprint models, but also shows outstanding training and data efficiency.

## VII. ACKNOWLEDGMENTS

This research was supported by the DOE Office of Science through the National Virtual Biotechnology Laboratory, a consortium of DOE national laboratories focused on response to COVID-19, with funding provided by the Coronavirus CARES Act and as part of the CANDLE project by the DOE-Exascale Computing Project (17-SC-20-SC). This research used the theory and computation resources of the Center for Functional Nanomaterials, which is a U.S. DOE Office

TABLE II  
MSEs OF THE DOCKING SCORE PREDICTION ON THE TEST SETS FROM TWO CFP MODELS (ECFP AND MORGAN) AND THREE GRAPH-BASED MODELS (GATEDGCN, GRAPHSAGE, AND MPNN). THE BEST PERFORMER FOR EACH TARGET IS INDICATED IN BOLD FONT. THE BASELINE MODEL CORRESPONDS TO THE NAIVE VARIATIONS BASED ON THE AVERAGE DOCKING SCORE IN THE TRAINING SETS.

Target	Baseline	ECFP	Morgan	GatedGCN	GraphSAGE	MPNN
3CLPro_pocket1	1.830	1.132	1.115	<b>1.031</b>	1.096	1.061
ADRP-ADPR_pocket1	1.588	0.237	0.202	<b>0.151</b>	0.157	<b>0.151</b>
ADRP-ADPR_pocket5	1.584	0.237	0.200	<b>0.149</b>	0.156	0.153
ADRP_pocket1	0.544	0.115	0.109	<b>0.085</b>	0.092	0.087
ADRP_pocket12	0.544	0.116	0.109	<b>0.085</b>	0.091	0.086
ADRP_pocket13	1.048	0.163	0.144	<b>0.104</b>	0.111	<b>0.104</b>
COV_pocket1	0.270	0.076	0.070	<b>0.054</b>	0.058	0.058
COV_pocket2	0.271	0.075	0.069	<b>0.055</b>	0.058	0.057
COV_pocket8	0.872	0.178	0.162	<b>0.125</b>	0.133	0.127
COV_pocket10	1.166	0.172	0.163	<b>0.120</b>	0.124	0.124
NSP9_pocket2	1.139	0.205	0.211	<b>0.164</b>	0.170	0.167
NSP9_pocket7	0.987	0.135	0.126	<b>0.089</b>	0.093	0.092
NSP15_pocket1	0.843	0.169	0.156	<b>0.122</b>	0.126	<b>0.122</b>
ORF7A_pocket2	0.397	0.149	0.144	<b>0.123</b>	0.127	0.124
PLPro_chainA_pocket3	0.566	0.167	0.161	<b>0.134</b>	0.142	0.135
PLPro_chainA_pocket23	0.927	0.253	0.242	0.199	0.207	<b>0.195</b>
PLPro_pocket6	0.843	0.157	0.139	0.112	0.118	<b>0.110</b>
PLPro_pocket50	1.335	0.293	0.272	<b>0.211</b>	0.222	0.215

of Science Facility, and the Scientific Data and Computing Center, a component of the Computational Science Initiative, at Brookhaven National Laboratory under contract no. DE-SC0012704. M.R.C. acknowledges the support by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship under Award Number DE-FG02-97ER25308. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under contract DE-AC02-06CH11357. W.C. would like to thank Yue Qian and Mark S. Hybertsen for helpful discussions.

## REFERENCES

- [1] S. F. Sousa, P. A. Fernandes, and M. J. Ramos, "Protein–ligand docking: current status and future challenges," *Proteins: Structure, Function, and Bioinformatics*, vol. 65, no. 1, pp. 15–26, 2006.
- [2] E. Fischer, "Einfluss der configuration auf die wirkung der enzyme," *Berichte der deutschen chemischen Gesellschaft*, vol. 27, no. 3, pp. 2985–2993, 1894.
- [3] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin, "A geometric approach to macromolecule–ligand interactions," *Journal of Molecular Biology*, vol. 161, no. 2, pp. 269–288, 1982.
- [4] P. S. Goodford, "Drug design by the method of receptor fit," *Journal of Medicinal Chemistry*, vol. 27, pp. 557–564, 1984.
- [5] —, "A computational procedure for determining energetically favorable binding sites on biologically important macromolecules," *Journal of Medicinal Chemistry*, vol. 28, no. 7, pp. 849–857, 1985.
- [6] L. G. Ferreira, R. N. Dos Santos, G. Oliva, and A. D. Andricopulo, "Molecular docking and structure-based drug design strategies," *Molecules*, vol. 20, no. 7, pp. 13384–13421, 2015.
- [7] R. S. Bohacek, C. McMartin, and W. C. Guida, "The art and practice of structure-based drug design: A molecular modeling perspective," *Medicinal Research Reviews*, vol. 16, no. 1, pp. 3–50, 1996.
- [8] L. Ruddigkeit, R. van Deursen, L. C. Blum, and J.-L. Reymond, "Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17," *Journal of Chemical Information and Modeling*, vol. 52, no. 11, pp. 2864–2875, 2012.
- [9] T. Sterling and J. J. Irwin, "Zinc 15 – ligand discovery for everyone," *Journal of Chemical Information and Modeling*, vol. 55, no. 11, pp. 2324–2337, 2015.
- [10] "REAL compounds," accessed: March 9, 2022. [Online]. Available: <https://enamine.net/compound-collections/real-compounds>
- [11] M. A. Khamis, W. Gomaa, and W. F. Ahmed, "Machine learning in computational docking," *Artificial intelligence in medicine*, vol. 63, no. 3, pp. 135–152, 2015.
- [12] K. Crampon, A. Giorkallos, M. Deldossi, S. Baud, and L. A. Stefanel, "Machine-learning methods for ligand–protein molecular docking," *Drug discovery today*, 2021.
- [13] P. J. Ballester and J. B. O. Mitchell, "A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking," *Bioinformatics*, vol. 26, no. 9, pp. 1169–1175, Mar 2010.
- [14] H. M. Ashtawy and N. R. Mahapatra, "Machine-learning scoring functions for identifying native poses of ligands docked to known and novel proteins," *BMC Bioinformatics*, vol. 16, no. 6, p. S3, Apr 2015.
- [15] L. Bucinsky, D. Borthák, M. Gall, J. Matúška, V. Milata, M. Pitoňák, M. Štekláč, D. Végh, and D. Zajaček, "Machine learning prediction of 3clpro sars-cov-2 docking scores," *Computational Biology and Chemistry*, vol. 98, p. 107656, 2022.
- [16] Y. O. Adeshina, E. J. Deeds, and J. Karanicolas, "Machine learning classification can reduce false positives in structure-based virtual screening," *Proceedings of the National Academy of Sciences*, vol. 117, no. 31, pp. 18477–18488, 2020.
- [17] P. O. Fernandes, D. M. Martins, A. de Souza Bozzi, J. P. A. Martins, A. H. de Moraes, and V. G. Maltarollo, "Molecular insights on abl kinase activation using tree-based machine learning models and molecular docking," *Molecular Diversity*, vol. 25, no. 3, pp. 1301–1314, Aug 2021.
- [18] K.-Y. Hsin, S. Ghosh, and H. Kitano, "Combining machine learning systems and multiple docking simulation packages to improve docking prediction reliability for network pharmacology," *PLoS ONE*, vol. 8, no. 12, p. e83922, Dec 2013.
- [19] T. Chandak, J. P. Mayginnis, H. Mayes, and C. F. Wong, "Using machine learning to improve ensemble docking for drug discovery," *Proteins: Structure, Function, and Bioinformatics*, vol. 88, no. 10, pp. 1263–1270, 2020.
- [20] F. Gentile, V. Agrawal, M. Hsing, A.-T. Ton, F. Ban, U. Norinder, M. E. Gleave, and A. Cherkasov, "Deep docking: a deep learning platform for augmentation of structure based drug discovery," *ACS central science*, vol. 6, no. 6, pp. 939–949, 2020.
- [21] A. Ahmed, B. Mam, and R. Sowdhamini, "Deelig: A deep learning approach to predict protein–ligand binding affinity," *Bioinformatics and Biology Insights*, vol. 15, 2021.
- [22] C. Shen, M. Krenn, S. Eppel, and A. Aspuru-Guzik, "Deep molecular dreaming: Inverse machine learning for de-novo molecular design and interpretability with surjective representations," *Machine Learning: Science and Technology*, vol. 2, p. 03LT02, 2021.

- [23] A. T. McNutt, P. Francoeur, R. Aggarwal, T. Masuda, R. Meli, M. Ragoza, J. Sunseri, and D. R. Koes, "Gnina 1.0: molecular docking with deep learning," *Journal of Cheminformatics*, vol. 13, no. 1, p. 43, Jun 2021.
- [24] Q. Bai, S. Liu, Y. Tian, T. Xu, A. J. Banegas-Luna, H. Pérez-Sánchez, J. Huang, H. Liu, and X. Yao, "Application advances of deep learning methods for de novo drug design and molecular dynamics simulation," *WIREs Computational Molecular Science*, p. e1581, 2021.
- [25] J. A. Morrone, J. K. Weber, T. Huynh, H. Luo, and W. D. Cornell, "Combining docking pose rank and structure with deep learning improves protein-ligand binding mode prediction over a baseline docking approach," *Journal of Chemical Information and Modeling*, vol. 60, no. 9, pp. 4170–4179, 2020.
- [26] Z. Liu, Y. Li, L. Han, J. Li, J. Liu, Z. Zhao, W. Nie, Y. Liu, and R. Wang, "Pdb-wide collection of binding data: current status of the pdbbind database," *Bioinformatics*, vol. 31, no. 3, pp. 405–412, Oct 2014.
- [27] L. Hu, M. L. Benson, R. D. Smith, M. G. Lerner, and H. A. Carlson, "Binding moad (mother of all databases)," *Proteins: Structure, Function, and Bioinformatics*, vol. 60, no. 3, pp. 333–340, 2005.
- [28] M. J. Hartshorn, M. L. Verdonk, G. Chessari, S. C. Brewerton, W. T. M. Mooij, P. N. Mortenson, and C. W. Murray, "Diverse, high-quality test set for the validation of protein-ligand docking performance," *Journal of Medicinal Chemistry*, vol. 50, no. 4, pp. 726–741, 2007.
- [29] E. J. Bjerrum, "Machine learning optimization of cross docking accuracy," *Computational Biology and Chemistry*, vol. 62, pp. 133–144, 2016.
- [30] O. Trott and A. J. Olson, "Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *Journal of Computational Chemistry*, vol. 31, no. 2, pp. 455–461, 2010.
- [31] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor, "Development and validation of a genetic algorithm for flexible docking," *Journal of Molecular Biology*, vol. 267, no. 3, pp. 727–748, 1997.
- [32] "WHO coronavirus (COVID-19) dashboard," <https://covid19.who.int>.
- [33] K. Ita, "Coronavirus disease (covid-19): Current status and prospects for drug and vaccine development," *Archives of medical research*, vol. 52, pp. 15–24, 2020.
- [34] C. Wu, Y. Liu, Y. Yang, P. Zhang, W. Zhong, Y. Wang, Q. Wang, Y. Xu, M. Li, X. Li *et al.*, "Analysis of therapeutic targets for sars-cov-2 and discovery of potential drugs by computational methods," *Acta Pharmaceutica Sinica B*, vol. 10, no. 5, pp. 766–788, 2020.
- [35] G. Mariano, R. J. Farthing, S. L. Lale-Farjat, and J. R. Bergeron, "Structural characterization of sars-cov-2: Where we are, and where we need to be," *Frontiers in Molecular Biosciences*, vol. 7, p. 344, 2020.
- [36] M. C. Pillon, M. N. Frazier, L. B. Dillard, J. G. Williams, S. Kocaman, J. M. Krahn, L. Perera, C. K. Hayne, J. Gordon, Z. D. Stewart *et al.*, "Cryo-em structures of the sars-cov-2 endoribonuclease nsp15 reveal insight into nuclease specificity and dynamics," *Nature communications*, vol. 12, no. 1, p. 636, 2021.
- [37] D. E. Gordon, G. M. Jang, M. Bouhaddou, J. Xu, K. Obernier, K. M. White, M. J. O'Meara, V. V. Rezelj, J. Z. Guo, D. L. Swaney *et al.*, "A sars-cov-2 protein interaction map reveals targets for drug repurposing," *Nature*, vol. 583, no. 7816, pp. 459–468, 2020.
- [38] M. K. Roe, N. A. Junod, A. R. Young, D. C. Beachboard, and C. C. Stobart, "Targeting novel structural and functional features of coronavirus protease nsp5 (3clpro, mpro) in the age of covid-19," *Journal of General Virology*, vol. 102, no. 3, 2021.
- [39] K. Michalska, Y. Kim, R. Jedrzejczak, N. I. Maltseva, L. Stols, M. Endres, and A. Joachimiak, "Crystal structures of SARS-CoV-2 ADP-ribose phosphatase: from the apo form to ligand complexes," *IUCrJ*, vol. 7, no. 5, pp. 814–824, Sep 2020.
- [40] M. T. Khan, M. Irfan, H. Ahsan, A. Ahmed, A. C. Kaushik, A. S. Khan, S. Chinnasamy, A. Ali, and D.-Q. Wei, "Structures of sars-cov-2 rna-binding proteins and therapeutic targets," *Intervirology*, vol. 64, pp. 55–68, 2021.
- [41] P. Krafcikova, J. Silhan, R. Nencka, and E. Boura, "Structural analysis of the sars-cov-2 methyltransferase complex involved in rna cap creation bound to sinefungin," *Nature Communications*, vol. 11, p. 3717, 2020.
- [42] Y. Kim, R. Jedrzejczak, N. I. Maltseva, M. Wilamowski, M. Endres, A. Godzik, K. Michalska, and A. Joachimiak, "Crystal structure of nsp15 endoribonuclease nendou from sars-cov-2," *Protein Science*, vol. 29, no. 7, pp. 1596–1605, 2020.
- [43] Z. Cao, H. Xia, R. Rajsbaum, X. Xia, H. Wang, and P.-Y. Shi, "Ubiquitination of sars-cov-2 orf7a promotes antagonism of interferon response," *Cellular & Molecular Immunology*, vol. 18, p. 746–748, 2021.
- [44] W. Rut, Z. Lv, M. Zmudzinski, S. Patchett, D. Nayak, S. J. Snipas, F. El Oualid, T. T. Huang, M. Bekes, M. Drag, and S. K. Olsen, "Activity profiling and crystal structures of inhibitor-bound sars-cov-2 papain-like protease: A framework for anti-covid-19 drug design," *Science Advances*, vol. 6, no. 42, 2020.
- [45] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson, "Autodock4 and autodocktools4: Automated docking with selective receptor flexibility," *Journal of Computational Chemistry*, vol. 30, no. 16, pp. 2785–2791, 2009.
- [46] V. Le Guilloux, P. Schmidtke, and P. Tuffery, "Fpocket: An open source platform for ligand pocket detection," *BMC Bioinformatics*, vol. 10, no. 1, p. 168, Jun 2009.
- [47] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open babel: An open chemical toolbox," *Journal of Cheminformatics*, vol. 3, no. 1, p. 33, Oct 2011.
- [48] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson, "Autodock4 and autodocktools4: Automated docking with selective receptor flexibility," *Journal of Computational Chemistry*, vol. 30, no. 16, pp. 2785–2791, 2009.
- [49] R. Huey, G. M. Morris, A. J. Olson, and D. S. Goodsell, "A semiempirical free energy force field with charge-based desolvation," *Journal of Computational Chemistry*, vol. 28, no. 6, pp. 1145–1152, 2007.
- [50] A. D. Hill and P. J. Reilly, "Scoring functions for autodock," in *Glycoinformatics*. Springer, 2015, pp. 467–474.
- [51] H. L. Morgan, "The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service," *Journal of Chemical Documentation*, vol. 5, no. 2, pp. 107–113, 1965.
- [52] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, 2010.
- [53] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Proceedings of the 28th International Conference on Neural Information Processing Systems—Volume 2*, 2015, pp. 2224–2232.
- [54] "Rdkit: Open-source cheminformatics software." [Online]. Available: <https://www.rdkit.org>
- [55] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," *arXiv:1511.05493*, 2015.
- [56] X. Bresson and T. Laurent, "Residual gated graph convnets," *arXiv:1711.07553*, 2017.
- [57] Z. Ying, J. You, C. Morris, X. Ren, W. L. Hamilton, and J. Leskovec, "Hierarchical graph representation learning with differentiable pooling," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [58] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *arXiv:1706.02216*, 2017.
- [59] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1263–1272.
- [60] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, 1988.
- [61] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv:1609.02907*, 2016.
- [62] M. Wang, L. Yu, D. Zheng, Q. Gan, Y. Gai, Z. Ye, M. Li, J. Zhou, Q. Huang, C. Ma, Z. Huang, Q. Guo, H. Zhang, H. Lin, J. Zhao, J. Li, A. Smola, and Z. Zhang, "Deep Graph Library: Towards Efficient and Scalable Deep Learning on Graphs," *arXiv:1909.01315*, 2019.
- [63] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, "Molecular graph convolutions: moving beyond fingerprints," *Journal of Computer-Aided Molecular Design*, vol. 30, no. 8, pp. 595–608, 2016.
- [64] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.
- [65] H. Öztürk, A. Özgür, and E. Ozkirimli, "Deepdta: deep drug-target binding affinity prediction," *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, 2018.

# Supplemental Material for Transferable Graph Neural Fingerprint Models for Quick Response to Future Bio-Threats

Wei Chen,<sup>1,\*</sup> Yihui Ren,<sup>2,\*</sup> Ai Kagawa,<sup>2</sup> Matthew R. Carbone,<sup>2</sup> Samuel Yen-Chi Chen,<sup>2</sup> Xiaohui Qu,<sup>1</sup> Shinjae Yoo,<sup>2</sup> Austin Clyde,<sup>3</sup> Arvind Ramanathan,<sup>4</sup> Rick L. Stevens,<sup>5</sup> Hubertus J. J. van Dam,<sup>6,†</sup> and Deyu Lu<sup>1,‡</sup>

<sup>1</sup>Center for Functional Nanomaterials, Brookhaven National Laboratory, Upton, New York 11973

<sup>2</sup>Computational Science Initiative, Brookhaven National Laboratory, Upton, New York 11973

<sup>3</sup>Computational Science & Data Science and Learning Division,  
Argonne National Laboratory, Lemont, Illinois 60439

<sup>4</sup>Data Science and Learning Division, Argonne National Laboratory, Lemont, Illinois 60439

<sup>5</sup>Computing, Environment and Life Sciences, Argonne National Laboratory, Lemont, Illinois 60439

<sup>6</sup>Condensed Matter Physics and Materials Science,  
Brookhaven National Laboratory, Upton, New York 11973

## Appendix A: GCN and GraphSAGE Model architectures

The node and edge embedding procedures map from vectors in the feature dimensions to vectors in the embedding dimensions respectively for nodes and edges. The node and edge feature and embedding dimensions are shown in the following tables.

Node feature dimensions	[11, 7, 9, 8, 6, 3]
Node embedding dimensions	[15, 15, 15, 10, 10, 5]

Edge feature dimensions	[5, 5]
Edge embedding dimensions	[35, 35]

The output of the universal neural fingerprint layers is the input of the fully connected, multilayer perceptron (MLP). This MLP consists of two layers, and the number of nodes in each layer and activation functions are listed in the following tables. ReLU is Rectified Linear Unit.

Number of nodes from inputs to outputs	[70, 35, 1]
Activation functions from the first layer to last layer	[ReLU, None]

## Appendix B: Details on training the machine learning models

The docking dataset was randomly split into 240,000, 30,000, and 30,457 samples for training, validation, and testing, respectively, using the `numpy.random.shuffle` function [1]. The validation set was used to optimize hyperparameters including the numbers of hidden layers and neurons as well as perform early stopping (maximum 300 epochs) and model selection. The mean squared error (MSE) was used as the loss function. The models were

\* Contributed equally to this work

† [hvandam@bnl.gov](mailto:hvandam@bnl.gov)

‡ [dlu@bnl.gov](mailto:dlu@bnl.gov)

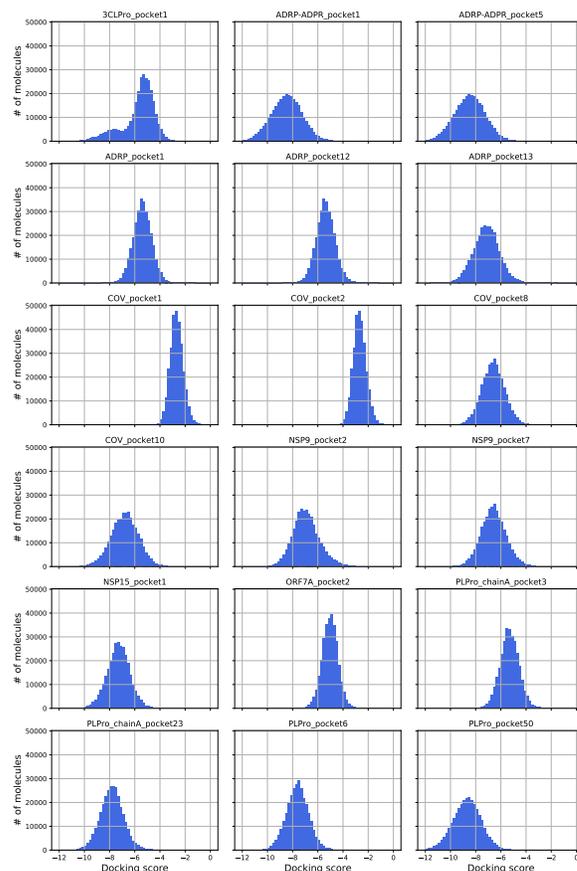


FIG. S1. Docking score distributions of molecules for 18 docking targets. The scores are in units of kcal/mol.

trained using the ADAM optimizer [2] with batch size 64. The initial learning rate was 0.001 with a reducing factor of 0.7 and a minimum learning rate of  $10^{-5}$ . This ML experiments were conducted using one node of the BNL institutional cluster. Our code used 20 Intel Xeon Gold 6248 processors and a NVIDIA V100 GPU on one node of the cluster. Each experiment for a particular target takes about 5-6 hours.

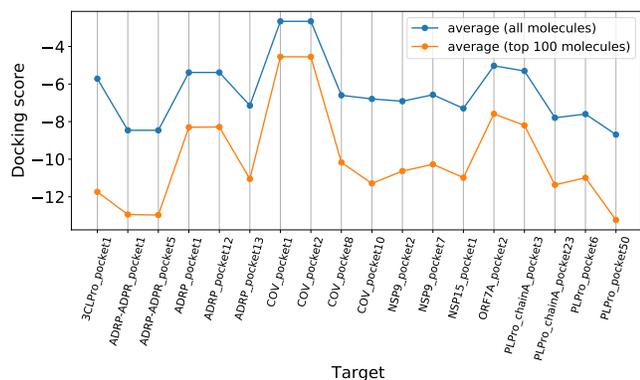


FIG. S2. Average docking scores of molecules for 18 docking targets.

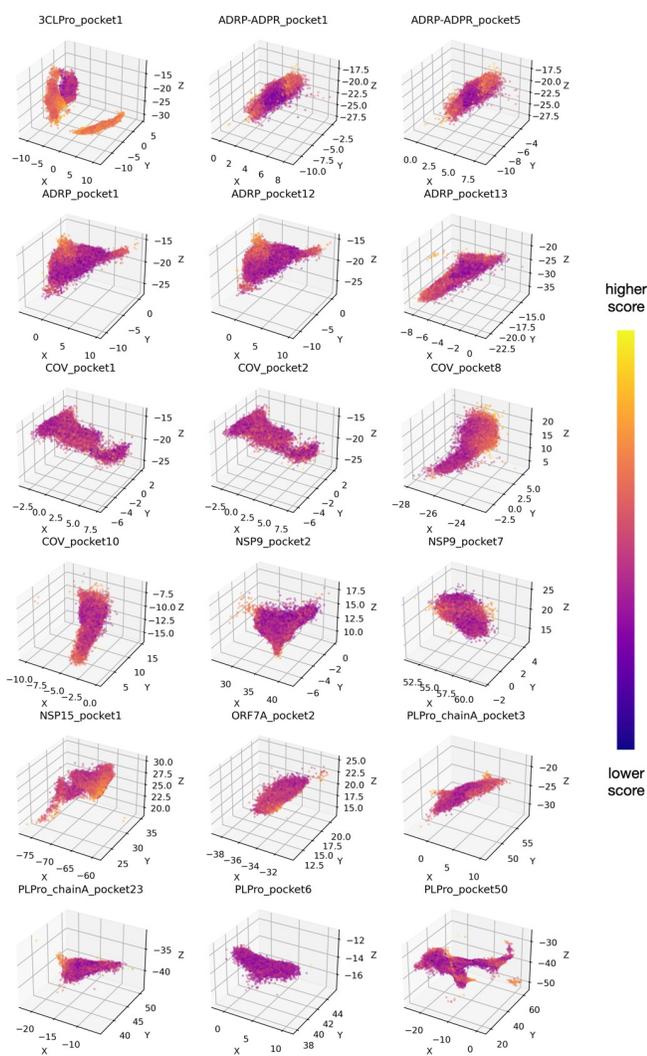


FIG. S3. Center-of-mass distributions of molecules for 18 docking targets. The coordinates are in units of angstrom.

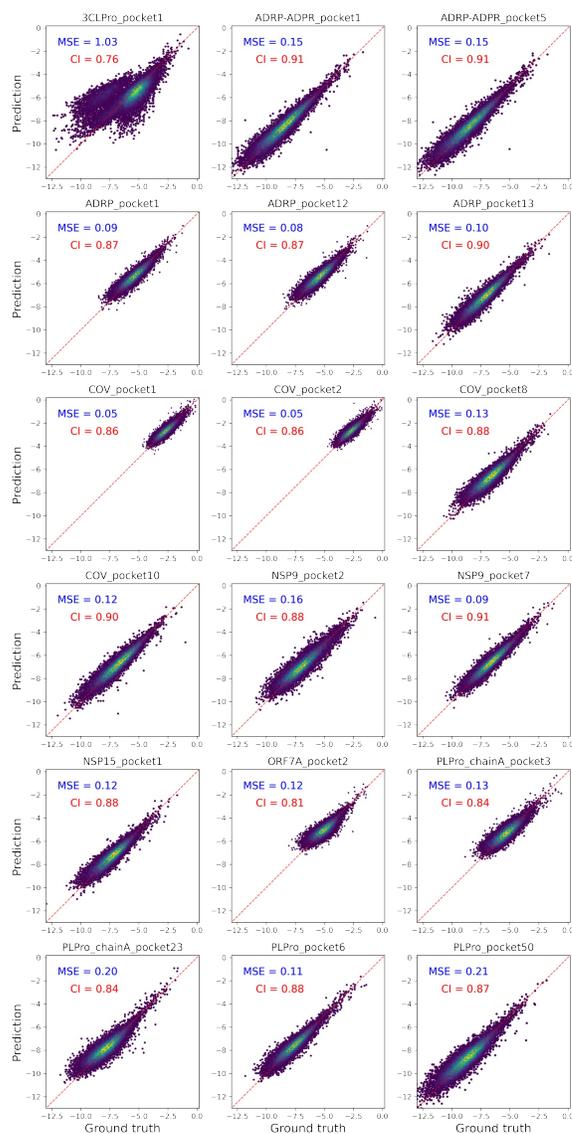


FIG. S4. Prediction performance on the test sets by the GatedGCN model. MSE stands for mean squared error and CI for concordance index.

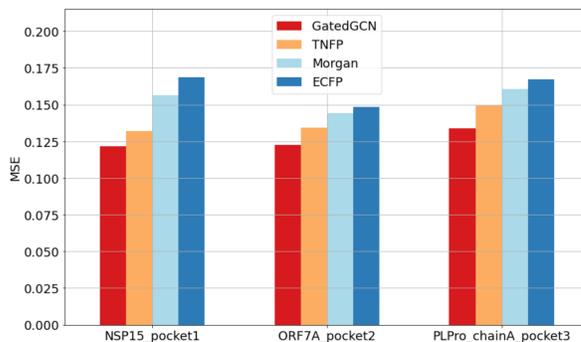


FIG. S5. Performance comparison of the TNFP model with the GatedGCN model (Fig. 2b) and two CFP models (Fig. 2a) tested on three targets.

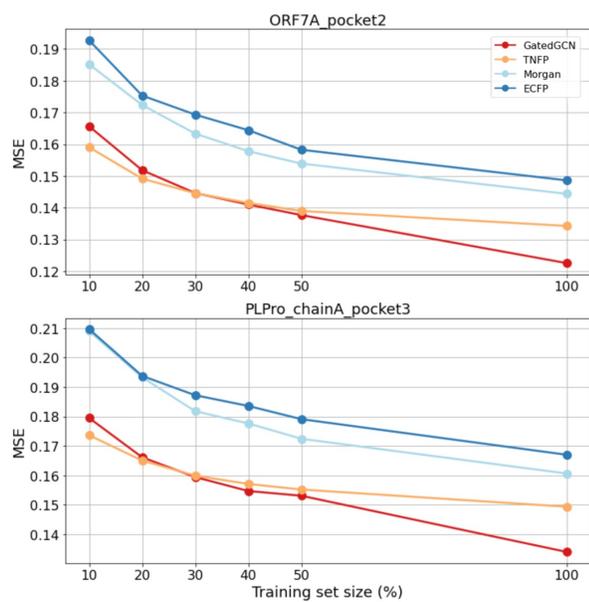


FIG. S6. Training data efficiency of four different models shown for two test targets.

- 
- [1] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, *et al.*, Nature **585**, 357 (2020).  
[2] D. P. Kingma and J. Ba, arXiv:1412.6980 (2014).