The further development of a graph convolutional neural network (GCNN) and methods for quantifying and systematically expanding its domain of applicability (DOA) demonstrated through ChemConX, which is a novel hands-on way to explore biochemical phenomena.

Benjamin M. Samudio[†], Yolanda Reyes[Ұ], Rafael Zamora-Resendiz[¥], and Silvia Crivelli[‡]

[†]Sierra College, 5100 Sierra College Boulevard, Rocklin, CA 95677
[Ұ]American River College, 4700 College Oak Drive, Sacramento, CA 95838
[¥] Hood College, 401 Rosemont Avenue, Frederick, MD 21701
[‡]Lawrence Berkeley National Lab, 1 Cyclotron Road, Berkeley CA 94720

1 Introduction: A review of our summer at LBNL in 2018

During the summer of 2018, three students from American River College and I were honored to be able to do research at Berkeley Lab in partnership with the Sustainable Horizons Institute. We set out to develop a neural network which could make crowdsourced medicine discovery faster, more informative, and increasingly accurate. We applied both 3D and graph convolutional neural networks to computationally evaluate (score) the interactions between medicine candidates (ligands) and the severe acute respiratory syndrome coronavirus (SARS-CoV) main protease ($3CL^{pro}$, receptor). Those interactions which score higher may be more likely to be good starting points for developing actual medicine. We chose to initially focus on $3CL^{pro}$ since it is crucial to the replication of SARS-CoV. This virus is a newly discovered etiological agent that was responsible for a deadly global outbreak which occurred in 2003 and 2004[1]. Though the outbreak has since been contained, it has the potential to reemerge and cause a severe threat to human health[2]. It is therefore imperative that we advance methods to proactively address the possibility of further outbreaks. What we learn may bolster efforts to combat similar viruses such as the Middle East respiratory syndrome coronavirus (MERS-CoV) which is currently endemic[3].

1.1 The curation of a ligand-receptor dataset and the development of a GCNN

In collaboration with Dr. Silvia Crivelli and Rafael Zamora-Resendiz, we formulated a chemical representation of ligand-receptor interactions which is amenable to deep learning, produced a dataset of over 2.8 million ligand-receptor structures and characterized their interactions, established that 3D convolutional neural networks are not ideal for our purposes, and in its place developed a GCNN (Figure 1). Physics-based computational methods such as molecular mechanics (MM) and quantum mechanics (QM) are used to virtually score medicine candidates. These methods typically involve a tradeoff between greater speed (MM is usually faster than QM) and increased accuracy (QM is usually more accurate than MM). The difference in speed between QM and MM can be 5 to 6 orders of magnitude. This makes the accurate scoring of large numbers of medicine candidates challenging. Our GCNN has been shown to score medicine candidates as fast as MM (about several milliseconds per evaluation) and has a high predictive power (with an R^2 of 0.85) on an MM dataset. It is anticipated that our GCNN will approach the accuracy of QM, when applied to the corresponding QM dataset, while maintaining its scoring speed. This paves the way for more accurate evaluations of larger numbers of medicine candidates thereby increasing the efficiency and pace of medicine discovery.

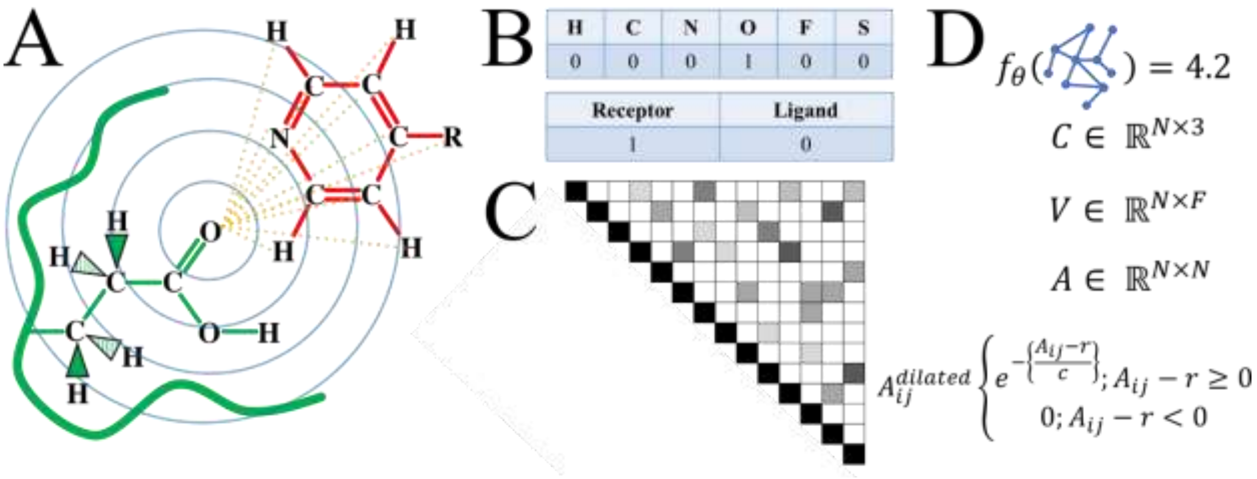
A
H
H
C
C
N
C
R
C
C
H
H
O
H
H
C
C
C
O
H
H
H
B
H C N O F S
0 0 0 1 0 0
Receptor Ligand
1 0
C
D
$f_\theta(\cdot) = 4.2$
$C \in \mathbb{R}^{N\times 3}$
$V \in \mathbb{R}^{N\times F}$
$A \in \mathbb{R}^{N\times N}$
$A_{ij}^{dilated} \begin{cases} e^{-\left\{\frac{A_{ij}-r}{c}\right\}}; A_{ij}-r \geq 0 \\ 0; A_{ij}-r < 0 \end{cases}$

Figure 1. Through a graph convolutional neural network (GCNN), the interactions between a small molecule ligand (panel A, red) and a receptor (panel A, green) can be represented by a graph with N nodes (atoms), atomic Cartesian coordinate tensor C, atomic feature (panel B) tensor V with F features, and adjacency matrices A (illustrated in panel C) and $A_{ij}^{dilated}$ *where c is a learned parameter and r is the radius of dilation (panel D).*

1.2 THE EXPLORATION OF METHODS TO DELINEATE THE GCNN'S DOA AND PRESENTATIONS AT INTERNATIONAL CONFERENCES

In addition to the curation of ligand-receptor structures and the development of the GCNN, we have explored methods to delineate the GCNN's DOA. It is desirable to qualify a neural network prediction with a confidence measure. Our GCNN is trained on ligand-receptor structures which span limited regions of physiochemical space. More reliable predictions may be made on those ligand-receptor structures which more closely (physiochemically) resemble the structures on which the GCNN was trained. We seek to establish a robust confidence measure which can help assess to which ligand-receptor structures our GCNN is most applicable and to serve as a guide in systematically exploring physiochemical space. Spectrophores™ may be thought of as a chemical "fingerprint" which can uniquely encode ligand-receptor structures and properties[4]. We have undertaken a preliminary exploration of Spectrophores as a proxy for our GCNN's DOA. Spectrophores can discern ligand-receptor structures with high resolution and hold promise as the basis for establishing a confidence measure applied to chemistry-based systems. I am pleased to report that we have presented our findings at the 2018 Supercomputing[5] and 2019 SIAM conferences[6]. These opportunities have been instrumental in allowing us to share our work, interact with fellow scientists, and find encouragement to continue in our scientific pursuits.

2 PROPOSAL: CURATING THE QM DATASET, QUANTIFYING AND EXPANDING DOAS, AND DEVELOPING CHEMCONX

We are excited at the prospect of continuing our work and collaboration during the summer of 2020. The timeline which we propose is detailed in Table 1.

Table 1. Proposed research timeline up to and including summer of 2020

Year 2020	Lead	Task
January	Ben Samudio	• Review research notes and reconnect with LBNL resources
	Yolanda Reyes	• Learn about research proposal and available resources
	Rafael Zamora-Resendiz	• Revisit chemistry-based graph convolutional neural network (GCNN)
February	Ben Samudio	• Curate a fragment library based on FDB-17
	Yolanda Reyes	• Connect wires to color sensors and interface them with a single-board microcontroller
	Rafael Zamora-Resendiz	• Familiarize Ben and Yolanda with the GCNN
March	Ben Samudio	• Dock fragments into SARS-CoV 3CL protease using Smina
	Yolanda Reyes	• Calibrate color sensors using plastic molecular model kits
	Rafael Zamora-Resendiz	• Help Ben and Yolanda find resources for running the GCNN
April	Ben Samudio	• Score fragment poses using DFT and QM:MM (1/3) • Determine pharmacophore positions on physical active site model • Investigate how the GCNN's domain of applicability (DOA) is formed
	Yolanda Reyes	• Join color sensors, microcontroller, and physical active site model into a prototype named ChemConX
May	Ben Samudio	• Score fragment poses using DFT and QM:MM (2/3) • Establish a method/proxy for quantifying the DOA
	Yolanda Reyes	• Develop the ChemConX algorithm (1/3)
June	Ben Samudio	• Score fragment poses using DFT and QM:MM (3/3) • Establish a method for systematically expanding the DOA
	Yolanda Reyes	• Develop the ChemConX algorithm (2/3)
July	Ben Samudio	• Translate pharmacophores into ChemConX algorithm weights
	Yolanda Reyes	• Develop the ChemConX algorithm (3/3)
August	Everyone	• Test ChemConX • Prepare to present research at conferences • Finalize publication draft and submit to ChemRxiv

2.1 Curating the QM dataset

The development of the MM dataset involved nearly 28 thousand individual simulations in which a ligand was virtually joined to the receptor (virtual docking) and several ligand-receptor structures (poses) were generated. These poses exist in a special part of the receptor, which resembles a pocket, called the active site. Each ligand-pair took about 1-minute wall clock time to simulate and resulted in about 100 poses. In total, about 2.8 million poses were generated and it took about 2 days wall clock time to complete. The NERSC supercomputers which made this possible provided a speedup of a factor of 10. Even with this speedup, however, attempting to develop the QM dataset on all 2.8 million poses, with QM calculations taking about 3 to 10 times longer than their MM counterparts, along with scheduling challenges, proved to be prohibitive.

We propose two methods to overcome the QM calculation challenges we face. The first is "smart" sampling in which pose Spectrophores are clustered into sets based on physiochemical diversity. For each ligand, only a small subset of the 100 poses, forming a diverse sample, need be calculated at the QM level thereby saving redundant computational effort. The second is to deconstruct the ligands into a set of fundamental fragments (Table 1, February). The number of fragments is less than the number of ligands since many ligands share fragments in common. These fragments would then be docked into the receptor at various centers within the active site (Figure 2). Poses would be sampled about these centers creating a new MM dataset (Table 1, March). QM calculations would then be performed to produce an associated QM dataset (Table 1, April). Since the fragments are smaller components of the ligands, the QM calculation times will be reduced. This will also provide a way to test to what extend the GCNN is able to extrapolate from fragments to whole ligand structures thereby providing a testbed for the quantification and expanding of the GCNN's DOA (Table 1, April-June).

2.2 QUANTIFYING AND EXPANDING DOAS

Sampling ligand fragment poses systematically throughout the receptor active site opens the opportunity to build on the benefits of using Spectrophores. Owing to the way Spectrophores are calculated, the position and orientation of poses in the active site are not explicitly accounted for in their unique "fingerprints". Since the active site is a complex physical and chemical landscape which can span tens of angstroms, it may be important to encode ligands in relation to the active site in the "fingerprints". We propose a method in which ligand position and orientation information is included with Spectrophores. For the purposes of this proposal, let's call these "enhanced fingerprints" Spectrophores$^+$. For each pose in the GCNN training set, Spectrophores$^+$ would be calculated and merged into an ensemble of Spectrophores$^+$. This would serve as a proxy of the GCNN's DOA since it encodes what the GCNN has been trained on. If the score for a new pose is to be predicted by the GCNN, then its individual Spectrophores$^+$ is calculated and compared to the ensemble Spectrophores$^+$. A percent match is computed between the two. If the match is above a threshold, then the prediction can be made with a certain level of confidence. If it is below this threshold, then full MM followed by QM calculations are performed and incorporated into the GCNN so that its DOA is expanded dynamically[7] (Table 1, April-June).

2.3 THE DEVELOPMENT OF CHEMCONX

ChemConX is a way to combine the predictive power of the GCNN with a hands-on approach to exploring biochemical phenomena and empowering crowdsourced medicine discovery. A papier-mache model of the 3CL^{pro} active site has been developed. This model is sized to be consistent with the scale of commodity hand-held molecular model kits[8]. The QM dataset and GCNN would be used to determine the effect on pose score of placing various types of ligand atoms at different locations within the active site. This would result in a 3D compositional "map" suggesting where particular atoms/chemical groups might be placed in order to improve the pose score. There would likely be "hotspots" (pharmacophores) within the active site where scores could be improved or worsened substantially. The map can be colored based on a scale of blue for improvements and red for deterioration of the score, analogous to a positive or negative sign, and the size of the hotspots can indicate the magnitude of these changes (Figure 3, panel A). The location of these hotspots would be translated to the papier-mache model and at those hotspots would be placed color sensors (Figure 3, panel B). The color sensors would then be connected to a microcontroller (Table 1, February-April) which contains an algorithm that pools signals from the color sensors (ChemConX algorithm). Participants would build medicine candidates from plastic molecular model kits using the papier-mache model as a structural guide. They would then connect the color sensors to various parts of their medicine candidates. The ChemConX algorithm would then determine which atom(s) are in proximity to which hotspots. The sign and magnitude of the hotspots, along with signals from the color sensors, are finally used to calculate a ChemConX score. This score is communicated to the participant in real time via an LCD screen which is attached to the microcontroller so that they may quickly evaluate and refine their medicine candidate.

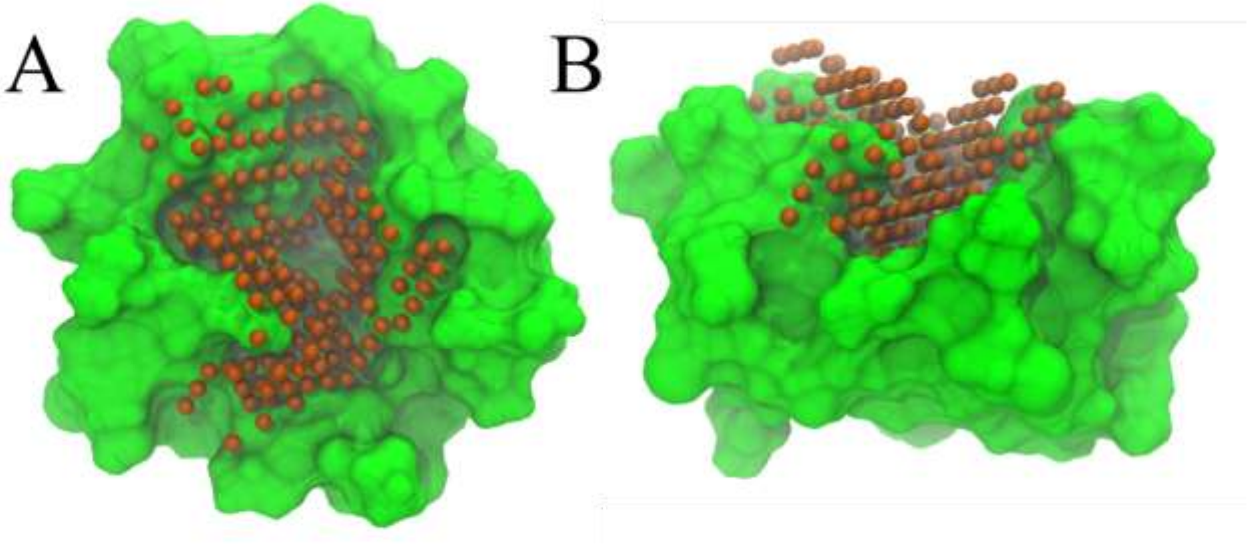

Figure 2. (A) Top and (B) side views of the SARS-CoV 3CL^{pro} active site. The surface of the active site is colored green. The orange spheres represent the center points of equally-sized cubes which enclose virtual docking search spaces.

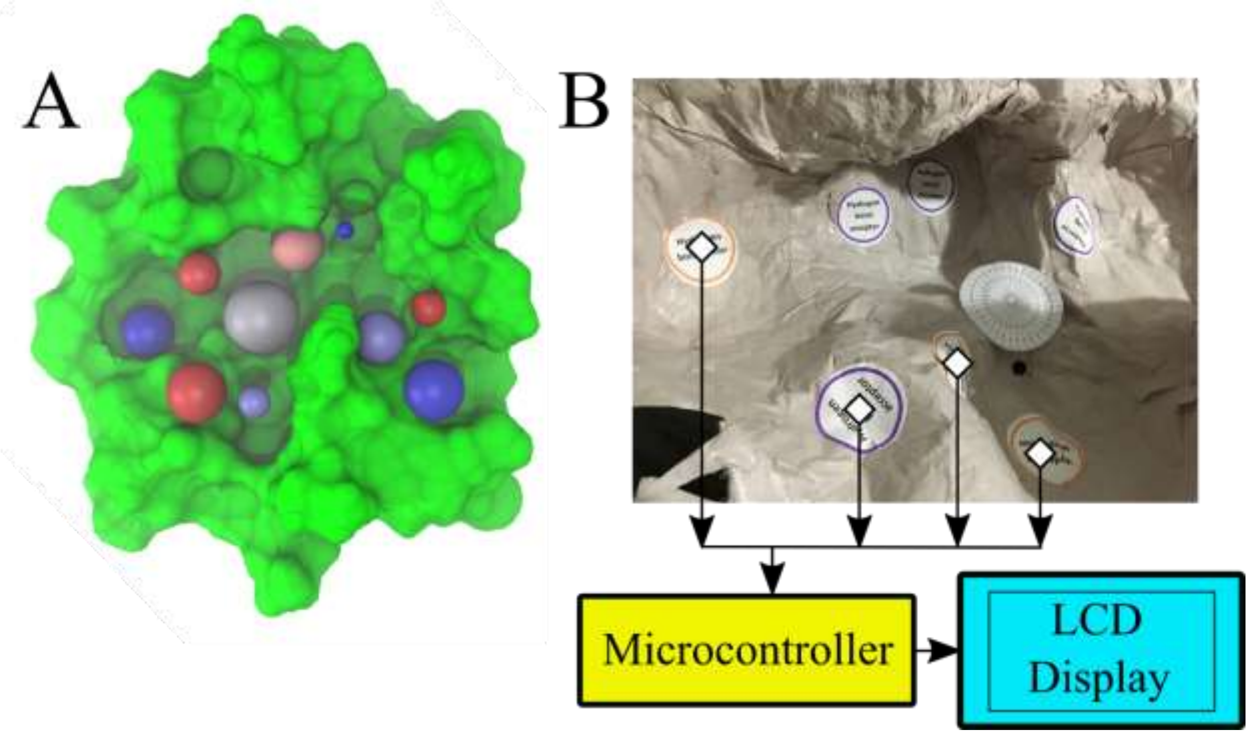

Figure 3. (A) GCNN-determined "hotspots", colored red and blue, within the active site of 3CL^{pro}, which is colored green. (B) The papier-mache model of the 3CL^{pro} active site. Superimposed on the image of this model are cartoon representations of color sensors (white diamonds), wires (black arrows) connecting to the microcontroller, and the LCD display.

3 REFERENCES

(1) SARS Basic Facts Sheet https://www.cdc.gov/sars/about/fs-sars.html (accessed Dec 27, 2017).

(2) Federal Select Agent Program https://www.selectagents.gov/SelectAgentsandToxinsList.html (accessed Dec 27, 2017).

(3) Everts, M.; Cihlar, T.; Bostwick, J. R.; Whitley, R. J. Accelerating Drug Development: Antiviral Therapies for Emerging Viruses as a Model. *Annu. Rev. Pharmacol. Toxicol.* **2017**, *57* (1), 155–169.

(4) Gladysz, R; Mendes Dos Santos, F; Langenaeker, W; Thijs, G; Augustyns, K; De Winter, H. 2018. Spectrophores as one-dimensional descriptors calculated from three-dimensional atomic properties: applications ranging from scaffold hopping to multi-target virtual screening. *J. Cheminform.* **2018**, *10* (9).

(5) Supercomputing 2018 Program https://sc18.supercomputing.org/proceedings/tech_poster/poster_files/post222s2-file3.pdf (accessed Jan 06, 2020).

(6) SIAM news https://sinews.siam.org/Details-Page/the-world-needs-new-and-better-medicines-now-1 (accessed Jan 06, 2020).

(7) Rupp, M.; Bauer, M. R.; Wilcken, R.; Lange, A.; Reutlinger, M.; Boeckler, F. M.; Schneider, G. Machine Learning Estimates of Natural Product Conformational Energies. *PLoS Comput. Biol.* **2014**, *10* (1).

(8) Samudio, B.; Karayi, G.; The development of sculptured molecular structure scale models at hands-on convenient size: A tool for the exploration, collaborative investigation, and communication of molecular phenomena. figshare. **2019** Preprint. https://doi.org/10.6084/m9.figshare.10247558.v2