OpeNNdd – Open Neural Networks for Drug Discovery

Crowdsourced medicine powered by people, machine learning, and supercomputing

Benjamin M. Samudio
Department of Chemistry
Sierra College
Rocklin, California, U.S.A.
SamudioBen@gmail.com

Shawn Shacterman, Bryce Kroencke, Nicholas Pavini
American River College
Sacramento, California, U.S.A.
OpeNNdd@gmail.com

Silvia Crivelli, Rafael Zamora-Resendiz
Lawrence Berkeley National Lab
Berkeley, California, U.S.A.
sncrivelli@lbl.gov

ABSTRACT

The current process required to bring new medicines to patients is prohibitively time-consuming and expensive leaving many diseases without therapeutics. In addition, new and reemerging diseases are increasing in prevalence across the globe at an alarming rate as current medications become ineffective at fighting them. The speed and scale with which new medicines are discovered must be increased if we are to effectively meet this challenge. OpeNNdd is a neural network platform which brings together people, machine learning, and supercomputing to solve the challenge of creating medicines. We have developed a novel neural network which quickly and accurately models candidate medicines interacting with a disease target, designed a metric to rigorously delineate and systematically expand its domain of applicability, and implemented a process that communicates the results of the neural network to participants in a readily interpretable way. OpeNNdd leverages the scale of supercomputing, the power and speed of neural networks, and the creativity of people across the globe in an open and collaborative way to protect and improve global health.

CCS CONCEPTS

• Crowdsourcing • Neural networks • Molecular simulation

KEYWORDS

Disease target, medicine candidate, pose, interaction energy/score, deep convolutional neural network, domain of applicability

1 Introduction

We empower citizen and career scientists alike to design ideas for medicine and share them with the world. The process is analogous to finding a key that fits just right into its cognate lock. Armed with the details of the lock's interior, the design of the key can be reverse engineered. Likewise, from the details of a disease target, candidate medicines may be constructed iteratively. The interactions between the target and candidate are optimized in a stepwise fashion yielding a candidate structure (pose) and total interaction energy/pose score at each step. Those candidates having the highest scores are likely to become good starting points for actual medicine.

Better Score = Better Starting Point

The interactions between the target and candidate are usually modeled using physics-based approaches, however, these typically involve a tradeoff between speed and accuracy making high-throughput evaluation of candidates challenging. We have developed a graph convolutional neural network (GCNN), based on a minimalistic chemical representation, which is able to predict these interactions both accurately and quickly while providing meaningful visual feedback. Our method, along with the capacity of supercomputing, pave the way for crowdsourced medicine discovery on a global scale. We first address the target 3CL protease ($3CL^{pro}$) of the severe acute respiratory syndrome related coronavirus (SARS-CoV). SARS-CoV caused a global outbreak in 2003 infecting over 8,000 people and killing nearly 800 of them [1]. SARS-CoV remains a threat and can reemerge at any time. In addition to the GCNN and visualization capabilities, we have developed a quantitative description of the domain of applicability (DOA) tailored to chemistry-centric GCNNs. We utilize a descriptor known as Spectrophores^{TM} [2] to delineate the DOA. This allows for a rigorous definition of prediction confidence limits, a systematic exploration of chemical space, and the progressive expansion of the GCNN training set.

DoA

2 COMPUTATIONAL DETAILS

2.1 The generation of GCNN input data

From the Zinc15 database [3], all protease-specific molecules (54,213) were filtered and standardized yielding 27,911 molecules. The $3CL^{pro}$ protein data bank (PDB) structure with accession code 5c5o was prepared for docking simulations. Smina [4] was used to generate approximately 100 poses per molecule resulting in 2.7 million unique poses (MM dataset). A GCNN was constructed in which a node feature tensor V (size N x 9) contains the one-hot encoding of the atom type and a coordinate feature tensor C (size N x 3) contains the Cartesian coordinates of each atom N.

Poses = P+L Combo

2.2 Spectrophore generation and testing

A spectrophore was created for each pose. An ensemble-pose spectrophore histogram (EH) is generated for a collection of poses (called the reference set) by first binning each descriptor value of each spectrophore, summing the values for each respective bin, and normalizing. A single-pose spectrophore histogram (SH) is created in a similar way. For each pose in a set (called the probe set), the SH belonging to that pose is compared to the EH of the refence set and a percent match is calculated. A histogram graph is then made of the resulting percent match values (Figure 2). An EH was created for 6950 poses (140 unique compounds). Six sets of SHs were created (probe sets), each with 1000 poses (20 unique compounds).

2.3 Data representation and the GCNN

All network models and operations were implemented in TensorFlow 1.5. The data were randomly split across poses into 70 % training, 10% validation, and 20% testing portions. An edge feature tensor A was generated by computing the Euclidean distance between points in C. An unordered convolution was performed according to feature tensors V and A. Batch normalization, softsign activation, and dropout were then applied in sequence. An attention matrix was used for node pooling. The final set of features were fed into a fully connected neural network layer of 1024 neurons with Relu activation and dropout.

2.4 Supercomputing and GPU resources

The generation and abstraction of chemical structures and creation of Spectrophores was carried out on the Cori and Edison supercomputers at Lawrence Berkeley National Lab. The final model was trained for a total of four epochs on a Quadro M4000 GPU. Training and testing required 1.5 and 0.25 hours per epoch respectively.

3 RESULTS

→ Reproducable?

3.1 Minimal chemical representation GCNN

A GCNN was trained on the MM dataset in which the label is the log of K_d which is similar to an interaction score (Table 1).

Table 1: Performance metrics for the MM dataset

Epoch	4
Training set loss	0.18
Test set loss	0.17
Training R^2	0.81
Testing R^2	0.82

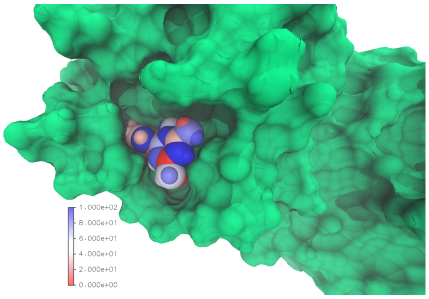

Figure 1: GCNN output visualization.

3.2 Delineating the domain of applicability

Spectrophores capture details of pose shape and chemical feature arrangement in 3D space and translates them into 1D molecular "fingerprints". We investigate whether the resolution of such "fingerprints" might be high enough to adequately classify poses into those that are similar to what the GCNN has been trained on (within the DOA) and those that are not (outside the DOA). As a proof-of-principle, a reference compound set, simulating the DOA of a GCNN, was constructed and compared to different versions of probe compound sets which were designed to be increasingly dissimilar to this reference (increasing probe index). Figure 2 clearly shows distinct distributions that move from higher to lower percent similarity as their respective compounds become less similar to the reference. This suggests that Spectrophores may be utilized to effectively delineate DOAs specific to chemistry.

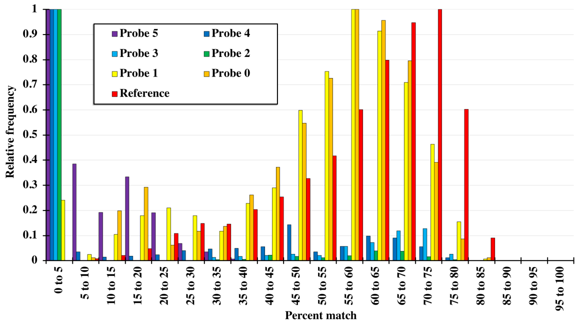

Figure 2: Spectrophore physiochemical resolution power.

3.3 The visual communication of GCNN results

The GCNN output is communicated to the participant visually. For a particular candidate, a prediction of the interaction energy is made. An atom is then removed and a new prediction is calculated. The difference between the two predictions quantifies that atom's contribution to the score. The atom is replaced and the process is repeated with the next atom. This is done in a serial way and across all atoms. The result is a heatmap representation [6] in which atoms are colored according to their contributions (Figure 1).

4 CONCLUSION

Towards realizing the goal of a fast, accurate, and informative medicine design capacity, we have: demonstrated that a GCNN based on a minimalistic chemical representation can produce models with high predictive power, created a dataset that promises to make such predictions highly accurate while being as fast as lower-levels of theory, described a method to delineate chemistry-centric DOAs, and showcased a medicine design-supportive visualization technique. Our results are made possible by a factor of 10 speedup afforded by supercomputers, intimating at the tremendous potential they have for advancing medicine.

ACKNOWLEDGMENTS

We would like to gratefully acknowledge the Sustainable Research Pathways (SRP) Program.

REFERENCES

[1] CDC. Centers for Disease Control and Prevention: SARS Basic Facts Sheet. Retrieved December 27, 2017 from http://www.cdc.gove/sars/about/fs-sars.html.

[2] Rafaela Gladysz, Fabio Mendes Dos Santos, Wilfried Langenaeker, Gert Thijs, Koen Augustyns, and Hans De Winter. 2018. Spectrophores as one-dimensional descriptors calculated from three-dimensional atomic properties: applications ranging from scaffold hopping to multi-target virtual screening. *J. Cheminform.* 10, 9 (March 2018), 24 pages. DOI: https://doi.org/10.1186/s13321-018-0268-9

[3] Teague Sterling and John J. Irwin. 2015. Zinc 15 – Ligand Discovery for Everyone. *J. Chem. Inf.* 55 (October 2015), 14 pages. DOI: https://doi.org/10.1021/acs.jcim.5b00559

[4] David Ryan Koes, Matthew P. Baumgartner, and Carlos J. Camacho. 2013. Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.* 53, 8 (February 2013), 12 pages. DOI: https://doi.org/10.1021/ci300604z

[5] Diola Bagayoko. 2014. Understanding density functional theory (DFT) and completing it in practice. *AIP Adv*. 4, 127104 (December 2014), 11 pages. DOI: https://doi.org/10.1063/1.4903408

[6] Matthew Ragoza, Joshua Hochuli, Elisa Idrobo, Jocelyn Sunseri, David Ryan Koes. 2017. Protein-Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model*. 57, 4 (April 2017), 16 pages. DOI: https://doi.org/10.1021/acs.jcim.6b00740