OpeNNdd: Open Neural Networks for Drug Discovery

Creating Free and Easy Methods for Designing Medicine

Bryce Kroenceke[cd], Shawn Shacterman[cd], Nicholas Pavini[cd], Ben Samudio[ad], Silvia Crivelli[b]

[a]Sierra College Department of Chemistry
[b]Lawrence Berkeley National Lab Computing Sciences
[c]American River College
[d]Fellow of Lawrence Berkeley National Lab

Introduction to OpeNNdd

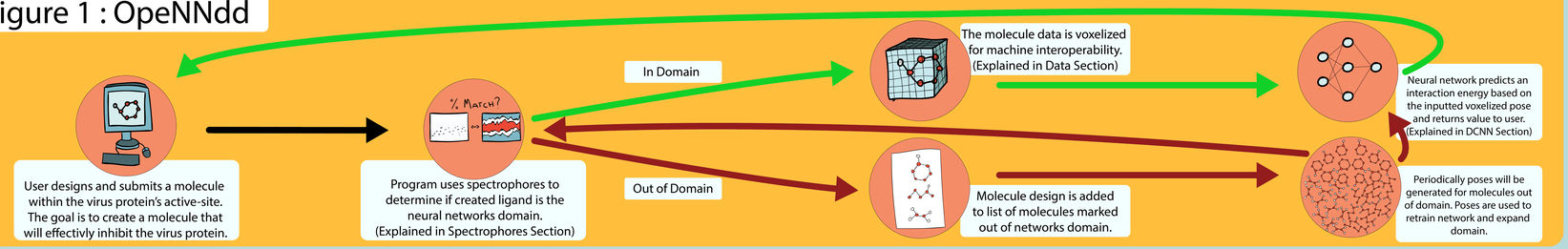
Problem

- Currently, designing medicine is an exclusive industry. This stems from the industry's proprietary software, requirement for high-level education, and expensive costs.
- The procedure for bringing a potential drug therapy to market, on average, costs around 2.6 billion dollars[1], takes 12 years[2], and has a failure rate of approximately 97%[3].

Solution

Create OpeNNdd (figure 1), an intuitive open-source program that gives students and citizen scientists the power to design medicine. Using High Performance Computing (HPC), the project has the resources needed to train deep convolutional neural networks to predict the quality of interactions occurring between a virus protein and user-engineered molecules.

Figure 1 : OpeNNdd



Dataset and Interoperability

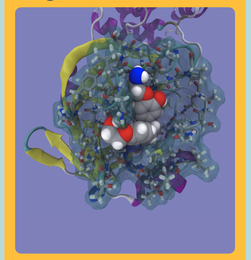
About the Dataset

HPC System Used for Data : Cori (Lawrence Berkeley National Lab)
Initial Protein : SARS coronavirus 3CL protease ($3CL^{pro}$) (figure 2)
Approximate Protein Dimensions : 25Å x 27Å x 28Å
Ligands : 27,912 ligands that have been tested against $3CL^{pro}$,
Dataset Structure :
- For each ligand, ~100 poses were created within $3CL^{pro}$,
- ~ 2.7 Million poses were generated in total.
- For each pose, the corresponding interaction energy was calculated.

Programs Used Within Cori to Generate Data :
- Molecular mechanical structures: Smina
- Quantum mechanical single point energies: Qchem
- Voxelized poses for interoperability: Custom python script (figure 3)

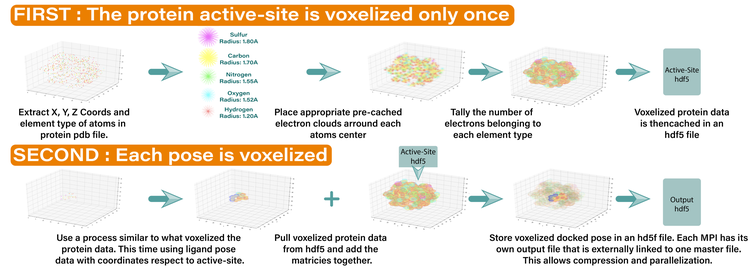
How HPC Was Used :
- Cori's nodes working in parallel decreased time to generate data.
  - Without parallelization : ~19.38 days. With parallelization : ~2 days.
- Cori hosted the datasets within the 20 TB user scratch.

Figure 2 : The Protein



Simulated image of student designed ligand posed inside of SARS 3CL Protease

Figure 3 : Voxelizing the Dataset for Interoperability



DCNNs to Find Interaction Energies

Why use Deep Learning?

What is Deep Learning Good at? :
Deep learning is advantageous when predicting on complex, high-dimensional data because it can establish a hierarchy of important structural features.
How OpeNNdd's Dataset Fits the Deep Learning Criteria :
- Protein dataset is complex and multi-dimensional.
- Chemical interactions can be found based on structural features.

Neural Network Info

Type of Network : 3D volumetric DCNN
HPC System Used : Summitdev (Oak Ridge National Lab)
Library : Tensorflow in Python
Channels : Six (H, C, N, O, F, and S electron counts)
Input : Voxelized protein/ligand poses
Output : Predicted interaction energy of inputted pose as a scalar value
Split Used : 70% Training, 10 % Validation, 20% Test
Model : Outlined in figure 4
Loss : Mean Squared Error (figure 6)
Optimizer : Adam Optimizer (enables an adaptive learning rate)
Batch Size : 25
How HPC Was Used :
- Summitdev's NVIDIA Tesla P100 GPUs are used for training models.
- Globus is used for fast data transfers between Cori and Summitdev.

Figure 5 : Placeholder of Network Results

Epochs	: 77
Mean Squared Error	: 8.36×10^{-2} Kd^2 on Test
Root Mean Squared Error	: 2.92×10^{-1} Kd on Test
Mean Absolute Percentage Error	: 3.45% On Test
Coefficient of Determination (R^2)	: 0.900

Figure 4 : Current Model

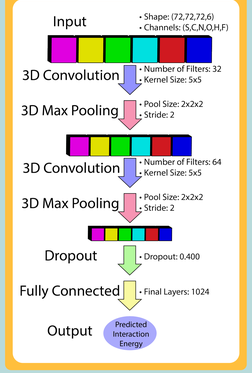
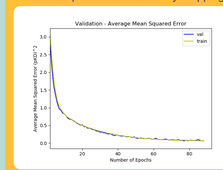


Figure 6 : Mean Squared Error and Early Stopping



Results from training on ~5000 Poses. Graph shows how overtraining is prevented with early stopping. The model stopped at 77 epochs right when overtraining would have began.

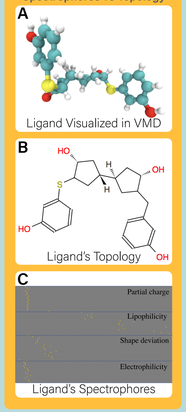
Spectrophores and Network Domain

Domain of Applicability in Chemistry Based Networks

Why is the Domain Important? :
- A neural network can only predict on data similar to the data it has been trained on.
- If the data used for testing is similar to the training data, it is considered to be in the domain.
- Predictions on objects within the domain should be more accurate than those outside of the domain

Current Industry Standard for Measuring the Domain of Chemistry Based Neural Networks :
2D topology of the 3D ligand is compared with the 2D topology of the neural networks training set to determine if the network is capable of making accurate predictions.

Figure 8: Visualizing Spectrophores vs Topology

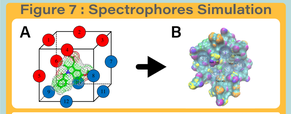


Spectrophores to Gauge Domain

What are Spectrophores? :
1D descriptors that accurately describe the 3D structures of molecules and their physicochemical properties. Spectrophores are generated with computer simulations where an artificial cage is placed around a molecule (figure 7A). The cage measures molecular properties in a process that is independent of position and orientation (figure 7B).

Why Spectrophores Over Topology? :
When using topology (figure 8B), a 2D description is being used to filter data that is occupying three dimensions. Comparatively, spectrophores (figure 8C) provide 3D descriptions that not only include structural information but physicochemical properties as well.

Figure 7 : Spectrophores Simulation

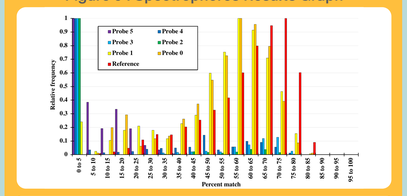


Spectrophores cage simulation(A) is used to find ligands property fields (B)

Experimenting Using Spectrophores

Process of Testing Spectrophores :
- Make a dataset where the spectrophores have been calculated for each pose.
- Normalize spectrophores values of the entire training set.
- Use spectrophores to bracket the test set by percent similarity to the training set (figure 9).
- Run the test set through the trained network and determine the correlation between percent match of spectrophores and the network's prediction accuracy.

Figure 9 : Spectrophores Results Graph



Graph illustrates spectrophores ability to delineate the domain of applicability

Conclusion and Acknowledgments

Accomplishments to Date

- Established flow of data through project.
- Generated an open-source protein ligand dataset.
- Developed protein voxelizer for interoperability.
- Created pipelines to efficiently handle the storage and retrieval of the large datasets.
- Tested practicality of spectrophores in chem based DCNNs
- Determined how to increase the domain of the neural network over time.
- Created 3D DCNNs with 4% error rates on scaled-down versions of the dataset.

Future Work

- Develop user interface and domain expansion system.
- Train models on the entirety of the protein database.
- Integrate functionality for more virus protein active-sites.
- Expand the platform into more schools and encourage public use.
- Further explore spectrophores as a method for screening the domain.
- Implement online crowdsourcing of designs.
- Widen the domain of DCNNs.

Citations

(1) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to Improve R&D Productivity: The Pharmaceutical industry's Grand Challenge. Nat. Rev. Drug Discov. 2010, 9, 203.
(2) DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs. J. Health Econ. 2016, 47, 20–33.
(3) Borhani, D. W.; Shaw, D. E. The Future of Molecular Dynamics Simulations in Drug Discovery. J. Comput. Aided. Mol. Des. 2012, 26 (1), 15–26.

Want to Contribute to OpeNNdd or Use Our Resources?

All aspects of OpeNNdd are open-source and accessible on Github.
- Code is well documented and written in a manner that is easy to read.
- Each program has a corresponding Wiki that was made to help familiarize users with functionality and how to use.
- Once more of the dataset has been created and curated, the database will be available for download.

Access Database Used : **LinkComingSoon**
Access Project Code : **https://github.com/Active-Site-Project/OpeNNdd.git**

Learn More About OpeNNdd : **http://www.studentsgivehope.org**
Questions or Suggestions Regarding the Project? : **Openndd@gmail.com**