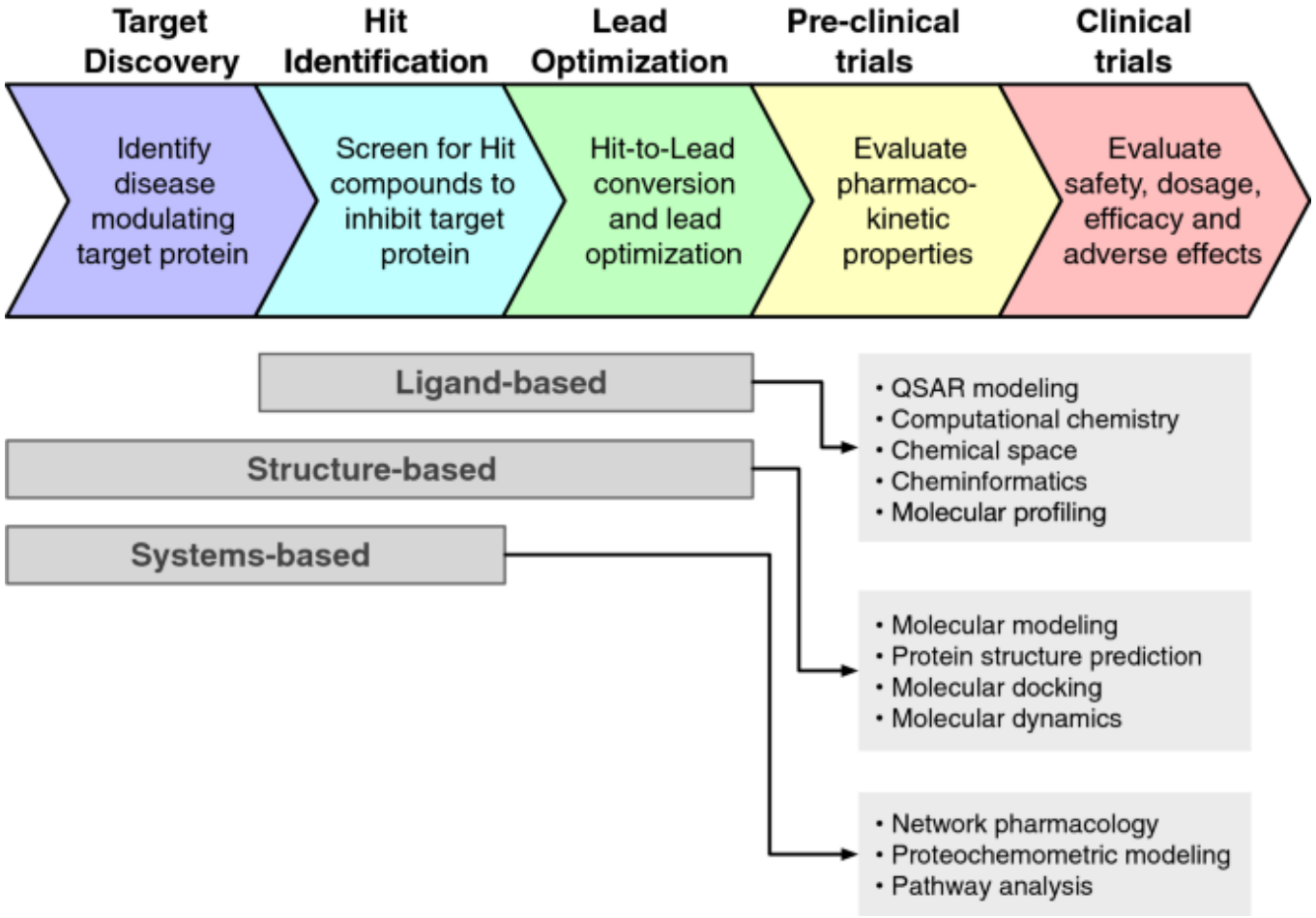
# Adapting Back-Translation for Theoretical Drug-Protein Scoring Mechanisms

Nathan Wood

July 30, 2020

# Table of Contents

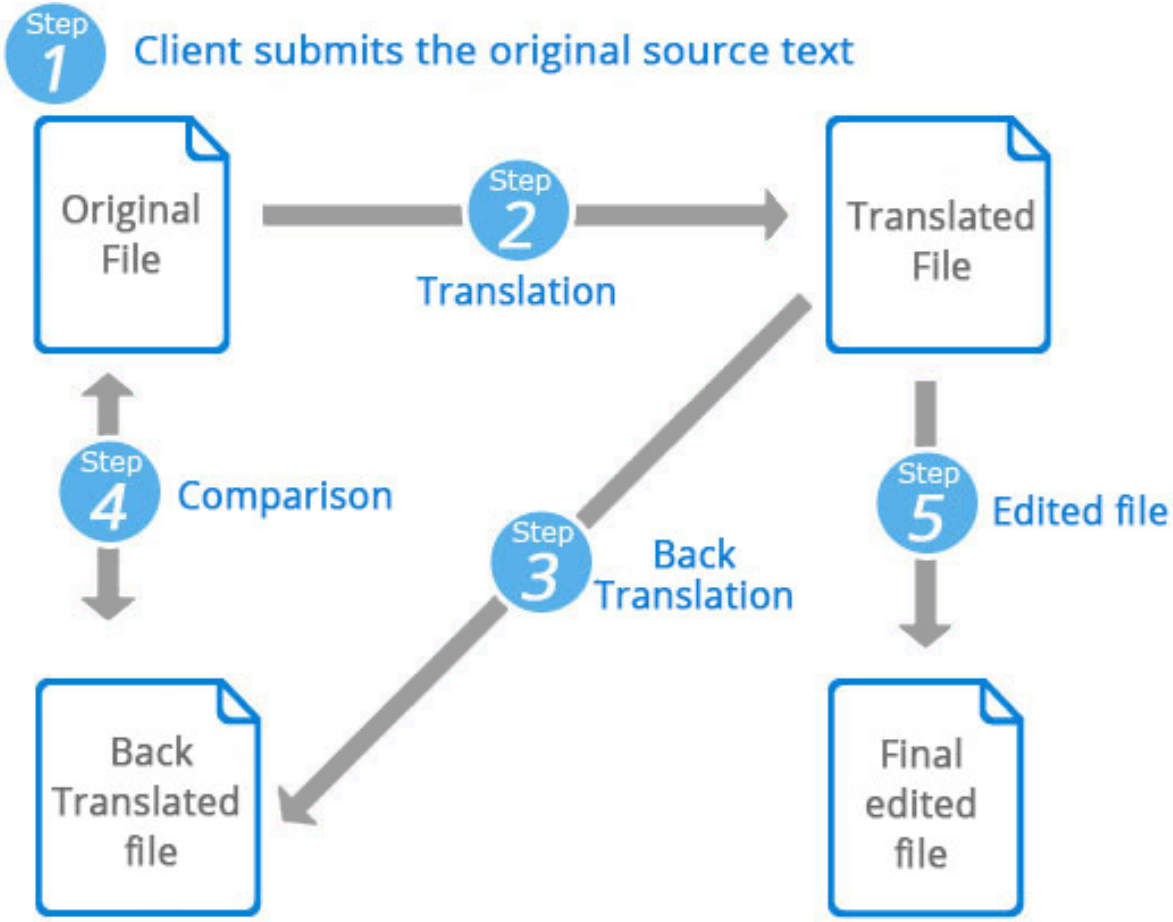
# Computational Drug Discovery Overview



# CDD Scoring Mechanisms (Koes et.al)

- Machine Learning methods are less computationally expensive than quantum-mechanical models
- Gnina
  - Divides structures in 24Å grid spaces to be processed using Convolutional Neural Networks
  - Rescoring poses generated using AutoDock Vina using CNN yields better results

# Computational Linguistics: Back-Translation



# Iterative Back-Translation (Hoang, et.al 2018)

- Back Translation typically improves model performance
- BLEU Score: specific sections are compared to a bitext reference, then averaged against the entire set

| Setting | French–English | | English–French | | Farsi–English | English-Farsi |
|---|---|---|---|---|---|---|
| | 100K | 1M | 100K | 1M | 100K | 100K |
| NMT baseline | 16.7 | 24.7 | 18.0 | 25.6 | 21.7 | 16.4 |
| back-translation | 22.1 | 27.8 | 21.5 | 27.0 | 22.1 | 16.7 |
| back-translation iterative+1 | 22.5 | - | 22.7 | - | 22.7 | 17.1 |
| back-translation iterative+2 | 22.6 | - | 22.6 | - | 22.6 | 17.2 |
| back-translation (w/ Moses) | 23.7 | 27.9 | 23.5 | 27.3 | 21.8 | 16.8 |

## Essential Question

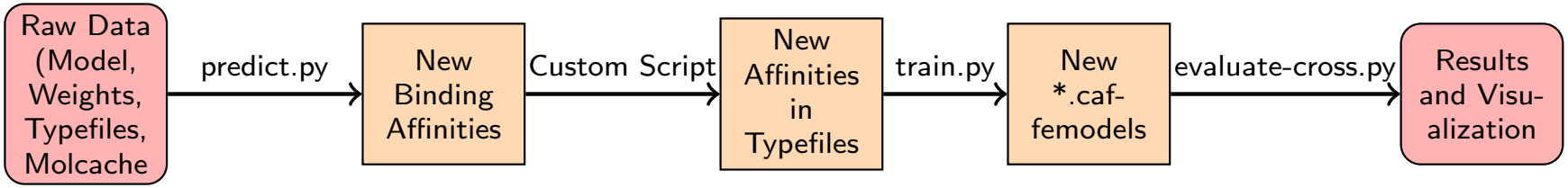
What if We Implement Unconfirmed Data into Our CNN Models?

# We Generate It !- A Proposal

- Like back-translation in neural-translation, adding computationally generated data may improve model performance
- This is accomplished by generating unconfirmed binding affinity data and protein-ligand docking poses

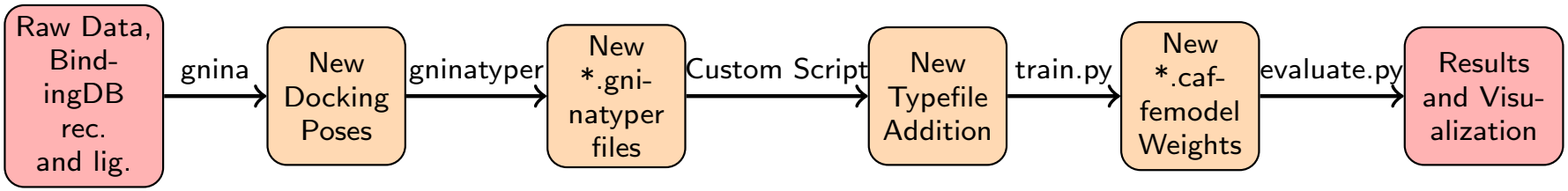
# Generated Binding Affinities

- Affinities are predicted from, and implemented into, the CrossDock 2020 Reduced Set
- Stochastic Gradient Descent: Seeds 0-4 are chosen
- new "*.caffemodel" weights are then cross-validated against the same set, but without the generated affinities
  - Root Mean Squared Error: maintaining a high binding affinity and low RMSE is optimal



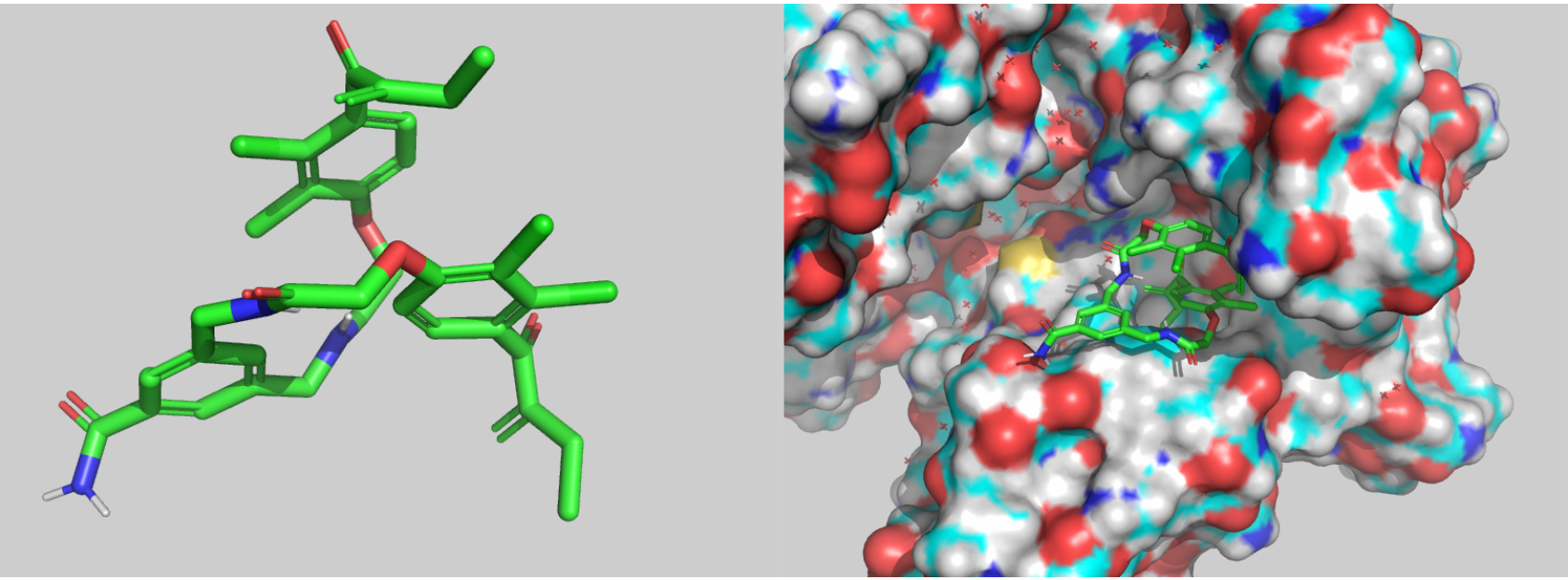
# Generated Protein-Ligand Docking Poses

- Docking Poses are generated from the BindingDB set, and implemented into the PDBBind 2016 set for training and evaluation
- Trained weights will be cross-validated against the BindDB Validation Set
  - Measured using Reciever Operating Characteristic Curve (true vs. false positives)
  - Root Mean Squared Deviation: 2Å(or less), and matching database is optimal

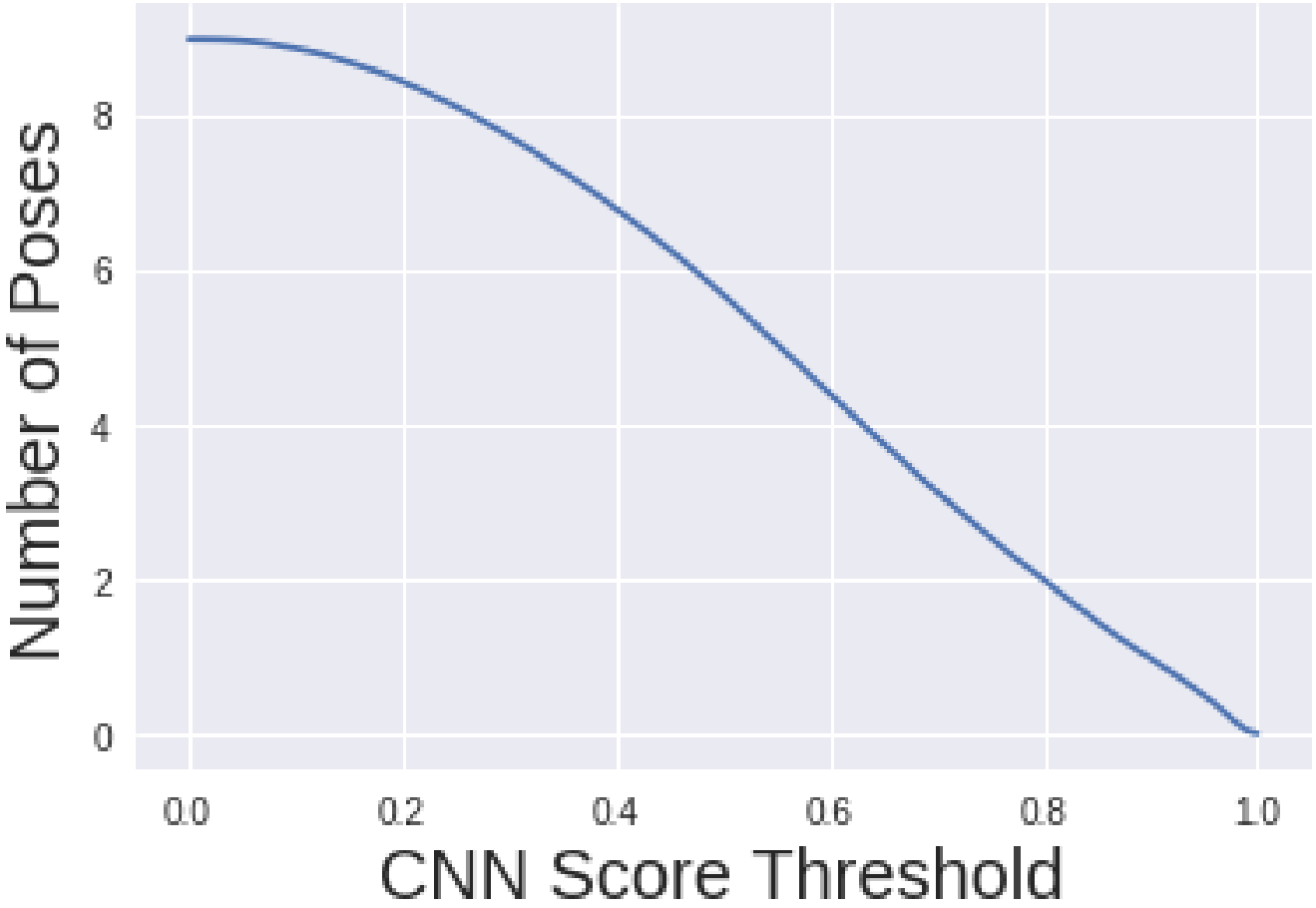


# Generated Docking Poses (Intermediate Results)

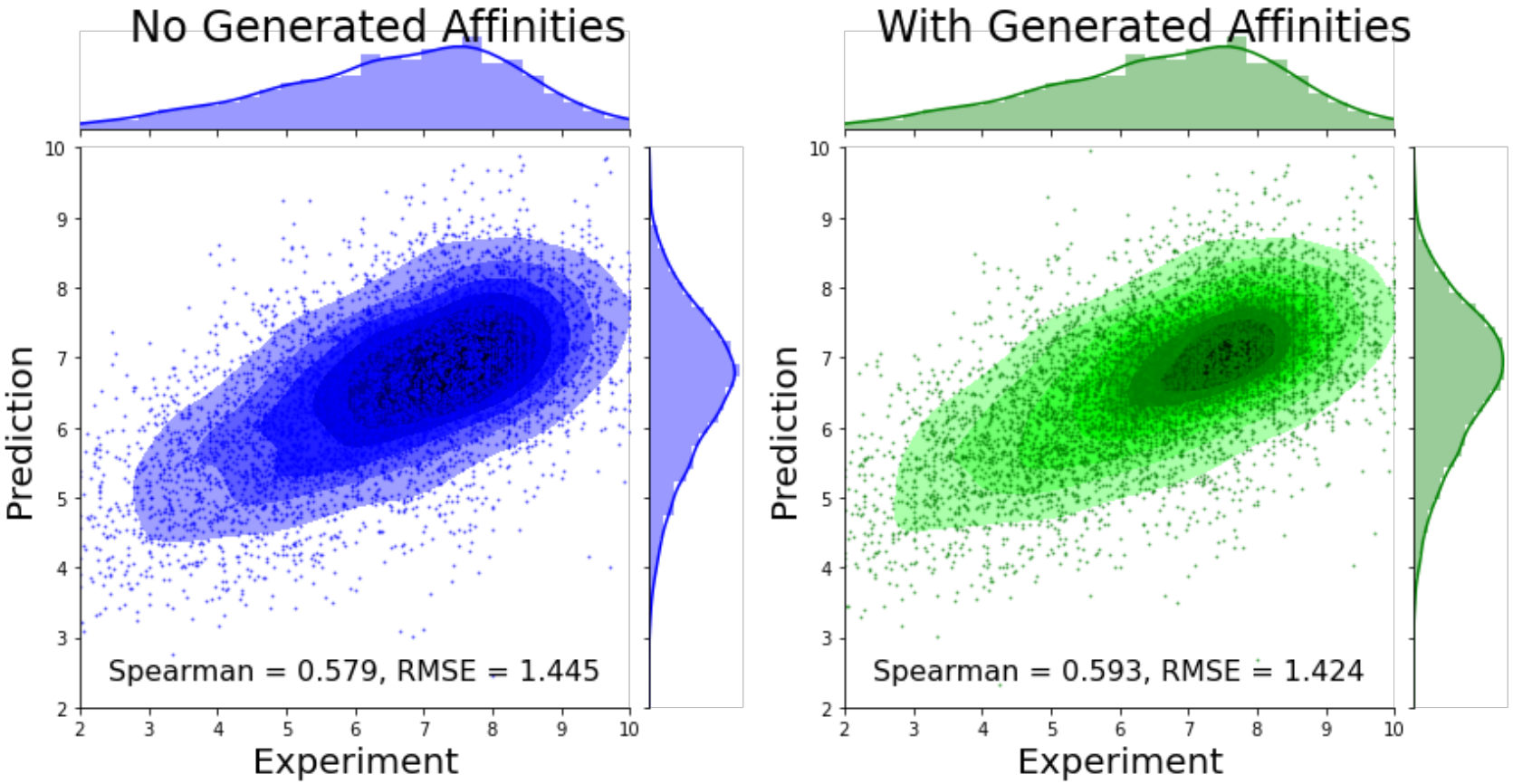
- 11gs (ligand 1) , CNN Score: .1515



- There is approximately 1 pose for every receptor (PDB) that has a CNN Score of .9 or higher



# Default 2018 IT2 Binding Affinities (Intermediate)



# Conclusions

- Adding generated binding affinities slightly improves model performance
  - Significance is unknown, highlighting a need to train an ensemble
- Assuming only one pose is correct, a .9 CNN Score threshold is preferable for labeling poses prior to model training

# Moving Forward

- Complete binding affinities workflow for CrossDock 2020 CompleteSet
- Complete binding affinities workflow for the DenseNet Model
- Generated Poses Set
  - Train, cluster, and evaluate
- Establish SMARTer goals
- Examine data more thoroughly prior to each workflow step

# Acknowledgements

- TECBio REU @ Pitt is supported by the National Science Foundation under Grant DBI-1659611
- Koes Group- Dr. David Koes, Paul Francoer, Johnathan King, Tomohide Masuda, Jocylyn Sunseri
- Dr. Ayoob, Adam Kohlhaas, and TECBIO 2020 Students

# References

Vu Cong Duy Hoang et al. "Iterative Back-Translation for Neural Machine Translation". In: *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*. Melbourne, Australia: Association for Computational Linguistics, July 2018. DOI: 10.18653/v1/W18-2703.

Matthew Ragoza et al. "Protein–Ligand Scoring with Convolutional Neural Networks". In: *Journal of Chemical Information and Modeling* 57.4 (2017). PMID: 28368587, pp. 942–957. DOI: 10.1021/acs.jcim.6b00740. eprint: https://doi.org/10.1021/acs.jcim.6b00740. URL: https://doi.org/10.1021/acs.jcim.6b00740.

# Questions and Answers

Questions