Adapting Back-Translation for Theoretical Drug-Protein Scoring Mechanisms

Nathan Wood, David Koes, Paul Francoer, Johnathan King, Daniel McNutt, Tomohide Masuda, Jocylyn Sunseri

Dept. of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15260

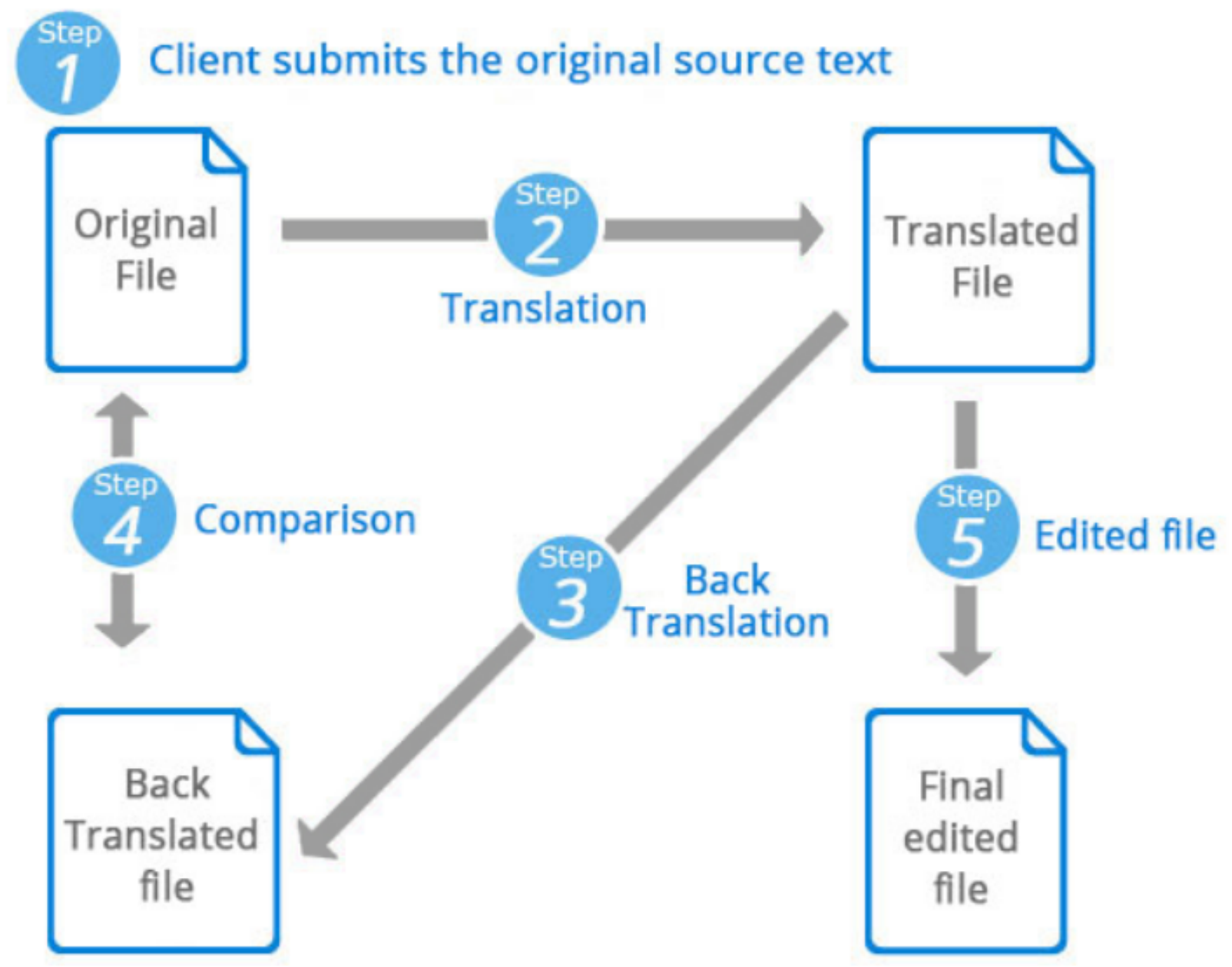
ABSTRACT

Back-translation was implemented into a Convolutional Neural Network based protein-ligand scoring mechanism in order to determine if adding unconfirmed and computationally generated data into the models will improve performance. This was accomplished by implementing generated binding affinities and docking poses into their respective datasets and retraining the CNN models. Despite time constraints and other setbacks, it was determined that generated binding affinities does improve model performance.

BACKGROUND

The use of Convolutional Neural Networks (CNN), a computational method typically associated with computer vision, has proven to be an effective means at scoring protein-ligand interactions. However, current CNN-based models typically have difficulty identifying crystal and near-crystal docking poses as high-scoring poses [2].

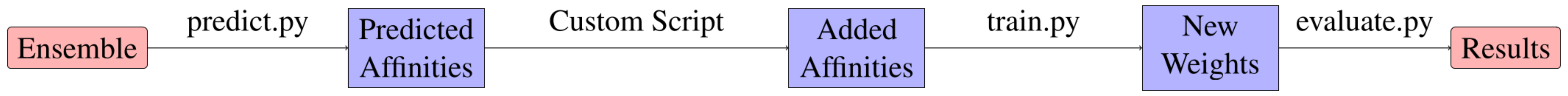
With respect to linguistics, neural translation is capable of implementing back-translation as an iterative process to evaluate and modify model performance, allowing monolingual data to be translated into another with a sparse data set [1]. These observations have lead to the idea that such principles can be applied to improve CNN model performance.



METHODOLOGY

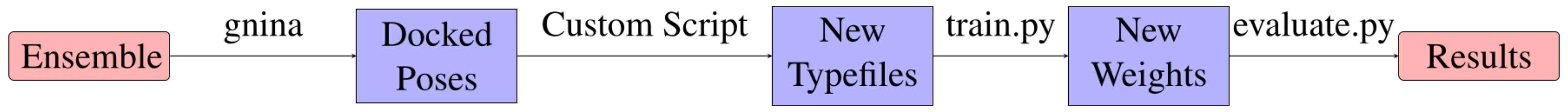
Generated Binding Affinities

Computationally generated binding affinities were run using the Default 2018 IT2 and DenseNet models against the reduced CrossDock 2020 dataset.


Ensemble
predict.py
Predicted Affinities
Custom Script
Added Affinities
train.py
New Weights
evaluate.py
Results

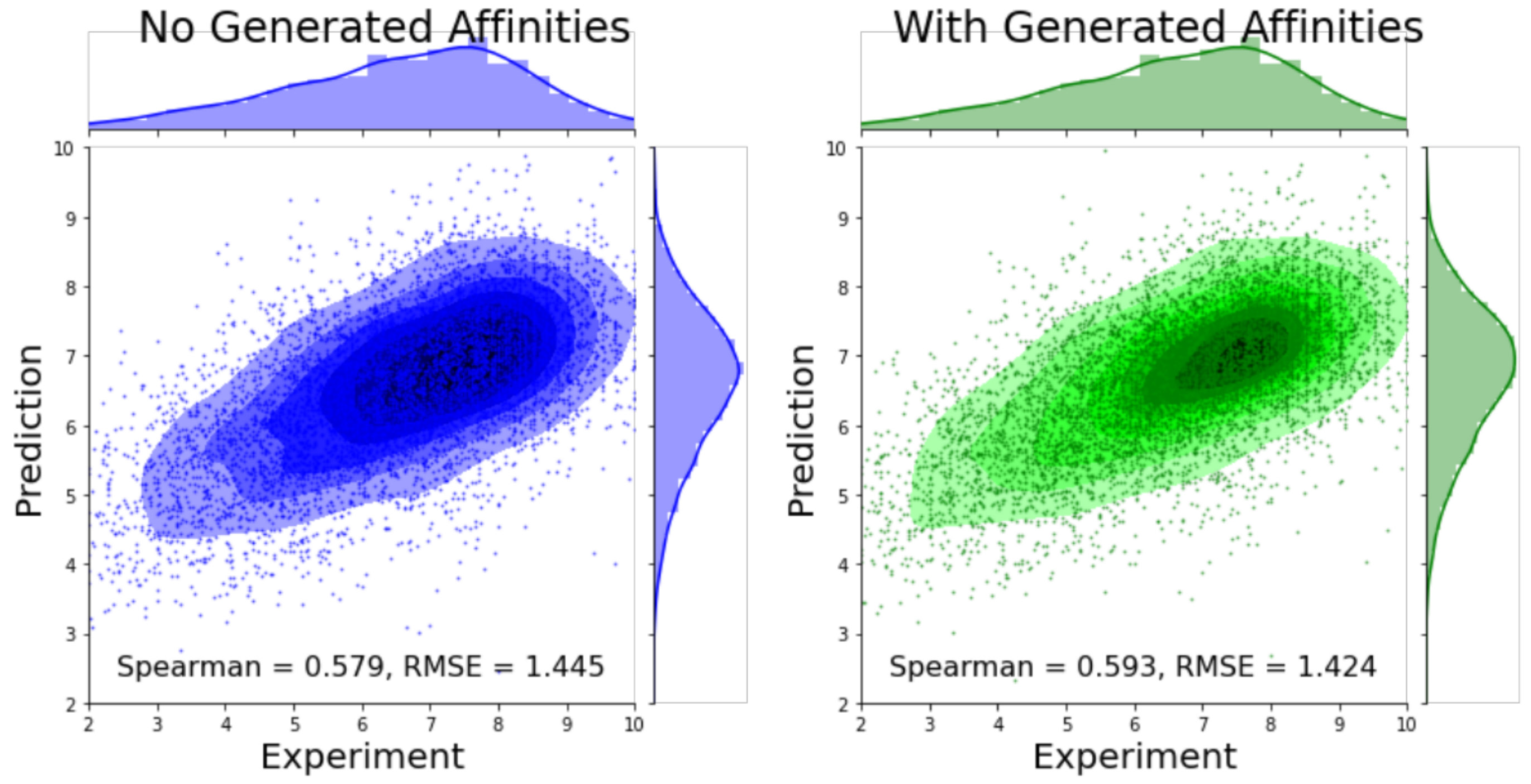
Generated Docking Poses

Computationally generated protein-ligand poses were generated using AutoDock Vina (through Gnina) from the Binding DB set, then implemented into the PDBBind 2016 set to be trained and evaluated.


Ensemble
gnina
Docked Poses
Custom Script
New Typefiles
train.py
New Weights
evaluate.py
Results

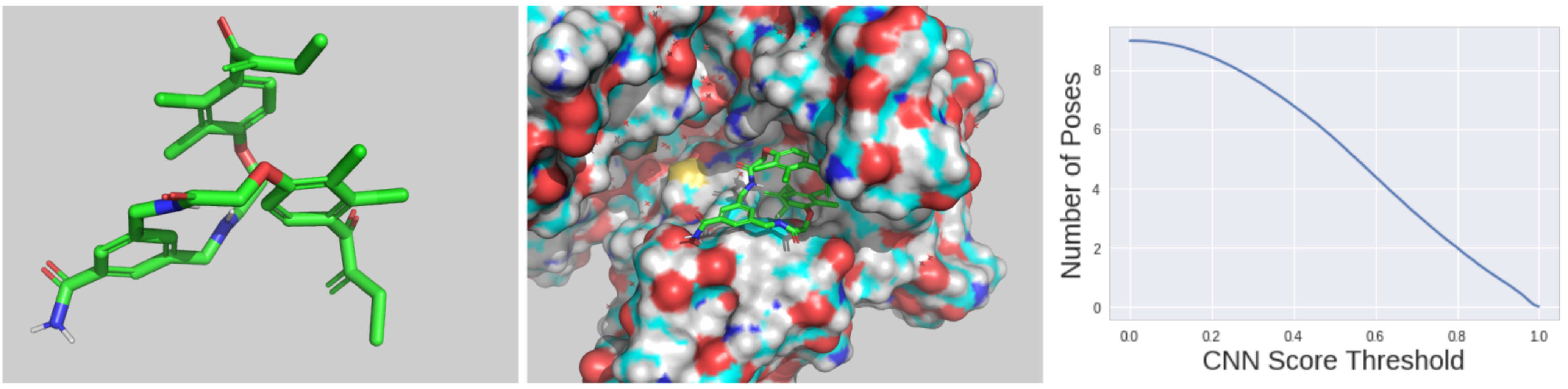
RESULTS

Generated Binding Affinities



Generated Docking Poses (Intermediate)

11GS02, a low scoring (.1515) protein ligand pose is an example of the poses generated by gnina using the BindDB set (left,center). Furthermore, a distribution of CNN scores was visualized (right), showing that there is approximately 1 poser per receptor that has a CNN Score of .9 or higher.



CONCLUSIONS

- Using generated affinities slightly improved model performance, exemplified by the RMSE of the added affinities set being lower than its counterpart (Default2018 IT2)
- Generated Poses with CNN Scores of .9 and higher should be labeled 1 when implemented into the PDBBind2016 models

FUTURE DIRECTIONS

This research serves as a potential starting point for:

- Implementing the same workflow on the CrossDock 2020 Complete Set, as well as the DenseNet model
- Continue to investigate the addition of computationally generated poses

ACKNOWLEDGEMENTS

TECBio REU @ Pitt is supported by the National Science Foundation under Grant DBI-1659611.

University of Pittsburgh TECBIO 2020 Staff and Students

Koes Group- Paul Francoer, Johnathan King, Dr. David Koes, Tomohide Masuda, Daniel McNutt, Jocylyn Sunseri

REFERENCES

[1] 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.

[2] 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.

LaTeX TikZposter