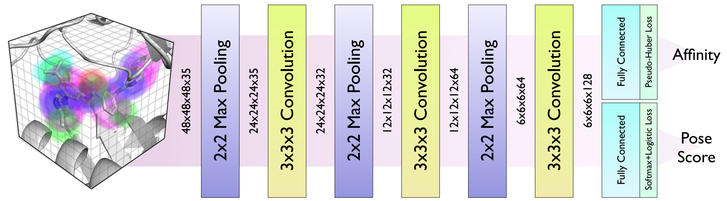
GNINADream: Deep Learning Delusions for Virtual Screening

Jocelyn Sunseri[1,2] and David Ryan Koes[2]
[1]Carnegie Mellon - University of Pittsburgh Joint PhD Program in Computational Biology,
[2]Department of Computational and Systems Biology University of Pittsburgh

Background

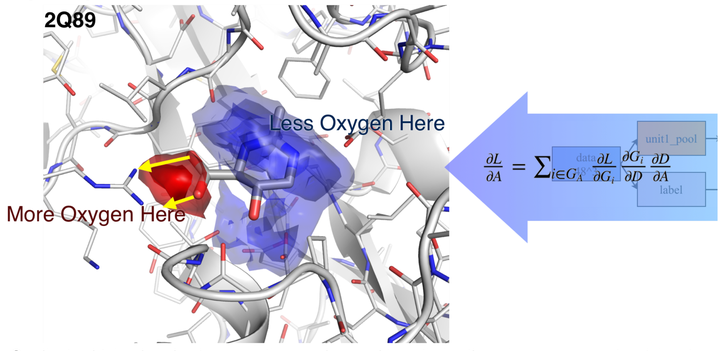
A central problem of drug discovery is understanding how small molecules bind to proteins. To identify promising new therapies, we need to accurately predict both *how* molecules bind to proteins, and *how strongly*. Going beyond the preliminary screening stage, we also need models of binding that enable us to predict how to optimize a lead molecule to make it a stronger and more specific binder for a chosen protein target. Our software, *gnina*, aims to address all these goals by leveraging convolutional neural networks (CNNs) to predict which molecules in a database bind a chosen protein strongly, how they bind, and to design molecules that bind even more strongly.

CNN Model Architecture



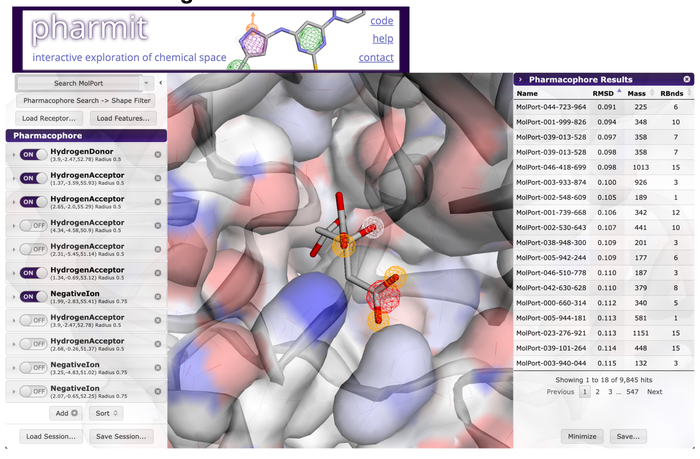
Via hyperparameter optimization we converged on this architecture, simultaneously optimizing for a binary classification (pose score) and regression (affinity) task. We have trained several variants with minor differences in architecture and training data. These are used to generate poses and predict binding strength; they can also be used to generate atom density maps that are applicable to tasks like virtual screening.

Input Optimization



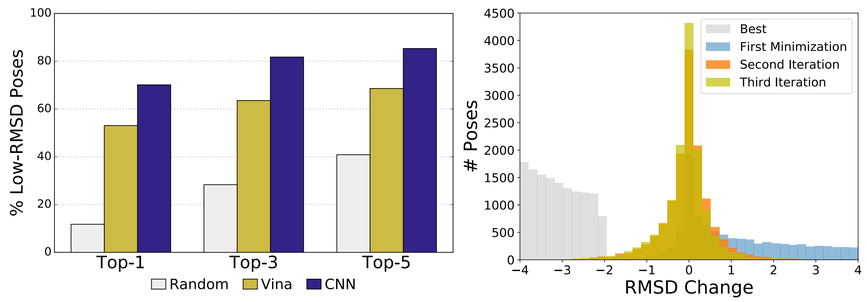
Gradients obtained via backpropagation can be used to optimize the original neural network input. While they can be directed onto the atoms used to generate the input grid to perform a conventional optimization of the ligand pose, they can also be used to modify the atomic density grids directly, producing novel molecules.

Virtual Screening

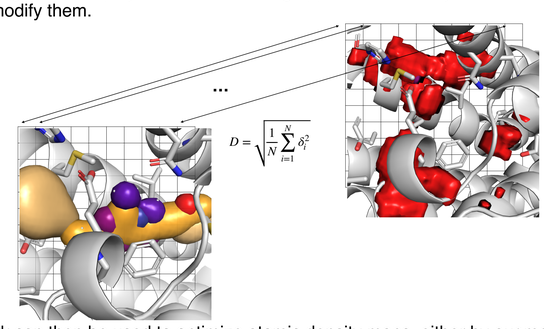


Methods

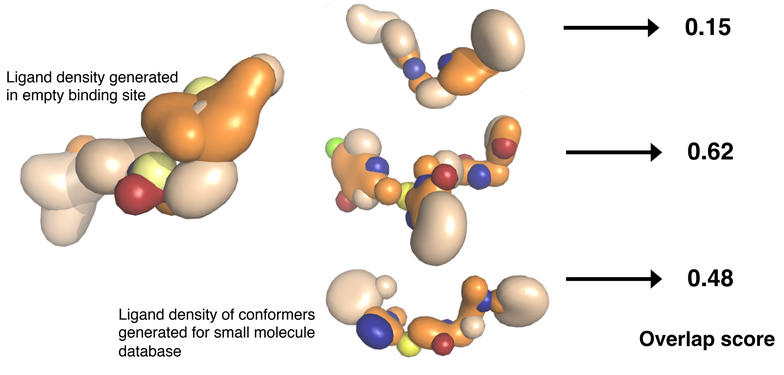
To use a CNN to perform a virtual screen or lead optimization, we first need to train a model that can discriminate binding modes from non-binding modes. Its performance depends significantly on training data, input representation, and training strategy.



(left) We train using clustered cross validation to provide a realistic estimate of generalization error. Docked poses are used as input, but crystal poses must be available so that distance from the crystal pose can be used as the determining factor in assigning the pose label. Data augmentation is used to provide regularization and simulated equivariance to input symmetries not preserved by our grid-based input representation. (right) We employ an iterative training approach to identify erroneous features learned by the model and modify them.



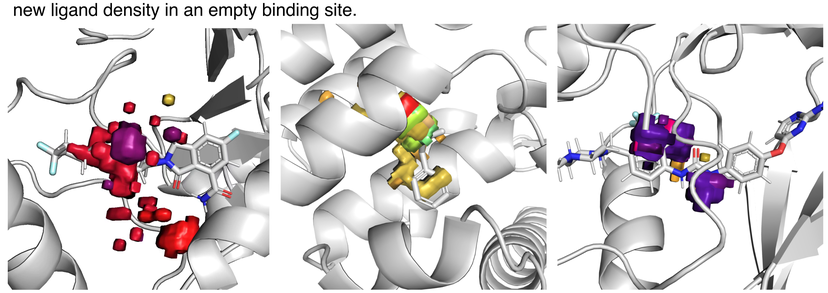
A trained network can then be used to optimize atomic density maps, either by augmenting existing maps or by performing de novo generation of the "optimal binder" for a given protein binding site. Alternatively, maps can be created by using other types of generative neural networks, such as generative adversarial networks. Then the similarity between the grids for real conformers and the generated densities must be computed.



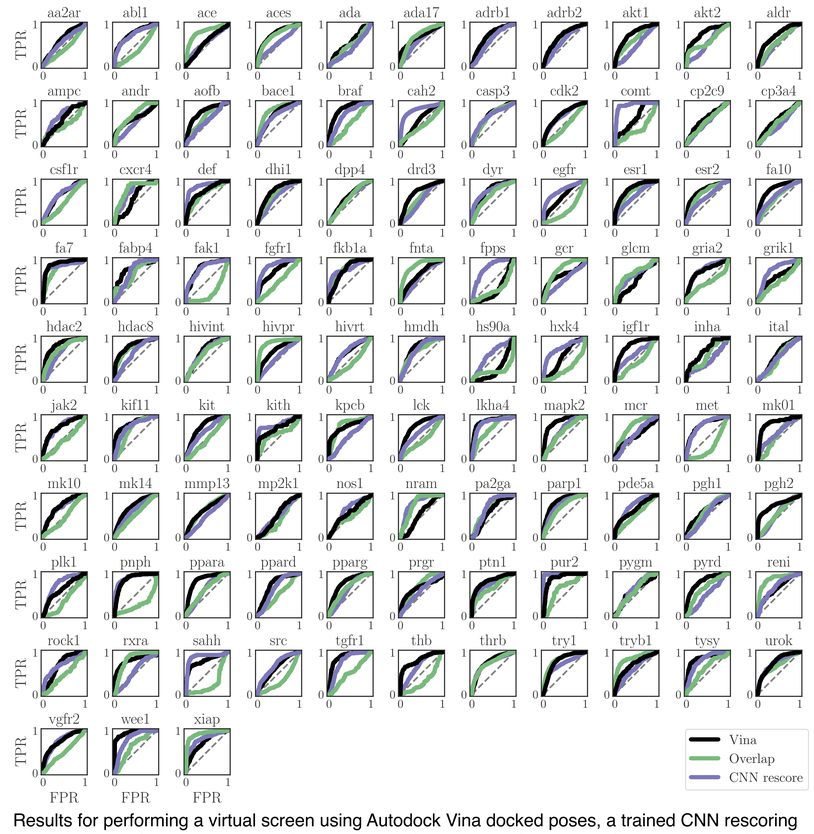
To generate density maps for ligands to be used in the virtual screen, docking may be performed; alternatively a pre-generated library of conformers (like that used in Pharmit, our pharmacophore-based virtual screening webserver) could be used to more quickly generate possible density maps. Once optimal and available ligand density maps have been generated, there are several possible methods for computing the similarity between the grids. These similarity scores (computed relative to the generated "optimal binder" map) are then used to rank the available compounds.

Results

Trained networks can be successfully used to augment existing ligand density or generate entirely new ligand density in an empty binding site.



In the above examples (PDB accession IDs 5A00, 184L, and 5A14), the crystal ligand is shown for reference, but it was not provided to the trained CNN as part of the input. Instead the receptor alone was provided, and the network produced the density shown in color.



Results for performing a virtual screen using Autodock Vina docked poses, a trained CNN rescoring the Vina-generated poses, or density maps generated from empty binding sites used to rank the docked poses using a Euclidean distance metric are shown above. Alternative training procedures and distance metrics are likely to improve performance.

Acknowledgements

This work is supported by a grant from Relay Therapeutics and R01GM108340 from the National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

References

Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., & Koes, D. R. (2017). Protein–Ligand scoring with Convolutional neural networks. *Journal of chemical information and modeling*, *57*(4), 942-957.

Sunseri, J., King, J. E., Francoeur, P. G., & Koes, D. R. (2018). Convolutional neural network scoring and minimization in the D3R 2017 community challenge. *Journal of computer-aided molecular design*, 1-16.