On the Application of Convolutional Neural Networks in the Prediction of Psychedelic Molecules

Nikolas I. D. Steckley[†]

[†]Washington State University

An Important Problem

There is increasing public and medical-professional interest being placed on psychedelics as medicine and tools of therapy in the treatment of PTSD, treatment-resistant depression, addiction, anxiety, and other mental health disorders [7]. Though the Federal Analogue Act had effectively closed the field of psychedelic chemistry since the late 70's, the models explored herein allow us to make progress in this important field once more using computational methods.

Herein we explore a model(s) that could be used to predict if a given molecule, specifically molecular structure, is ”psychedelic”. This model could be employed to seek out novel psychedelic drugs that do not have the side effects of presently available psychedelics (e.g. the hangover associated with MDMA) or that are honed for intended effects and treatment. Alternatively, this model could also be used for screening out psychedelically-active drugs when this is not a desired property. However, relatively few of such compounds are known, limiting the amount of data with which to train a deep learning model.

Definitions

For the purposes of machine learning, it is very difficult to give definition to a ”psychedelic molecule”. Psychedelics are generally thought of being comprised of two types: classical psychedelics - such as LSD, psilocybin, and DMT - are said to be serotonin-2A (5-HT2A) agonists while ”entacogens” such MDMA are said to be serotonin releasing. However, not all 5-HT2A agonists are psychedelic and there are counterexamples to each of these distinctions. Furthermore, the question of psychedelic activity it seems is more than just a question of effect on 5-HT2A alone: Ray found that each psychedelic compound has a distinct receptrome profile and that activity can not be determined by merely 5-HT2A activity alone. [3]

Theoretical Models

There are three models that show promise for further investigation. While all three models could theoretically answer the question of a candidate molecules' psychedelic-ability, each tries to answer this question in a different way and as such is different in either input and/or structure.

The first is simply a fully-connected neural network whose input is a flat array of chemical descriptors; we will refer to this model as the Simple Neural Net (SNN). This model assumes that the molecular descriptors contain all the information needed to determine psychedelic-ability and would rely more heavily on data (need more) than the other two.

The second and third models are similar to each other in that they both incorporate information from Ray's Receptorome data. In both of these models information about a candidate molecule's binding-affinities per nuereoreceptor is used in accordance with the receptorome data to determine how similar its activity is to known psychedelics. In the second model, the molecule's binding affinity at each receptor is determined using available standard metrics and algorithms. We refer to this as the Receptrome Basic (RB) model.

In the third model, a Convolutional Nueral Network (CNN) is employed to learn qualities of the molecule from its shape in a more individualized fashion. The CNN then provides an array of binding-affinities, possibly transformed if needed with a few fully-connected (FC) layers to map to Ray's standardized values, that is then passed on to the receptorome-trained layers. This model, dubbed CNN with Receptrome (CNNR) is the most promising for further research. Note that the first half of this network relies on a CNN for binding affinity prediction, while the second half is independently trained against receptrome data; to the researcher's knowledge, this is a novel form of network architecture and design in and of itself, whose technique warrants further study in its own right.

The advantage of using a CNN is that by design these networks learn information by ”building up” a picture. In this case: the CNN learns about the properties of the molecule by analyzing its constituent parts, the atoms, and how they are bonded together locally, and the way that subgroup is bonded to the neighboring subgroup and so on.

In particular, this CNN portion of the model could be made to provide binding affinities for each receptor all at once (see figure 1, path A) or alternatively CNNs could be called in sequence, each trained for a specific receptor (path B). In either case, a CNN based on Ragoza, et. al's Protein-Ligand Scoring with CNN's [2] or Öztürk, et. al's DeepDTA [1], would be the best places to start architecturally for the CNN, as these architectures have already been found to be effective for these types of purposes.

My intuition tells me that chemical descriptors alone, as in the SNN model, will be insufficient to accurately predict activity; if they were then this problem would be solved for a much larger class of drugs and better understood in general. I believe that the latter two models hold more promise because of their ability to learn about the molecule in a holistic manner, learning about and from integrated intricacies rather than discrete classifications and loosely-related data points.

No matter the model, using multitask learning will not hinder training performance if implemented correctly. Specifically, the model could be trained to identify not only if something is ”psychedelic” (perhaps more accurately within this context thought of as ”psychoactive”) but also if it is more ”classically psychedelic-leaning” or ”entactogenic-leaning”, assuming such labels can be prescribed to the training data. But this becomes difficult quickly as these are perhaps better thought of as two ends of a continuum rather than mutually exclusive choices; indeed, some psychedelics exhibit qualities of both types, e.g 2C-B.

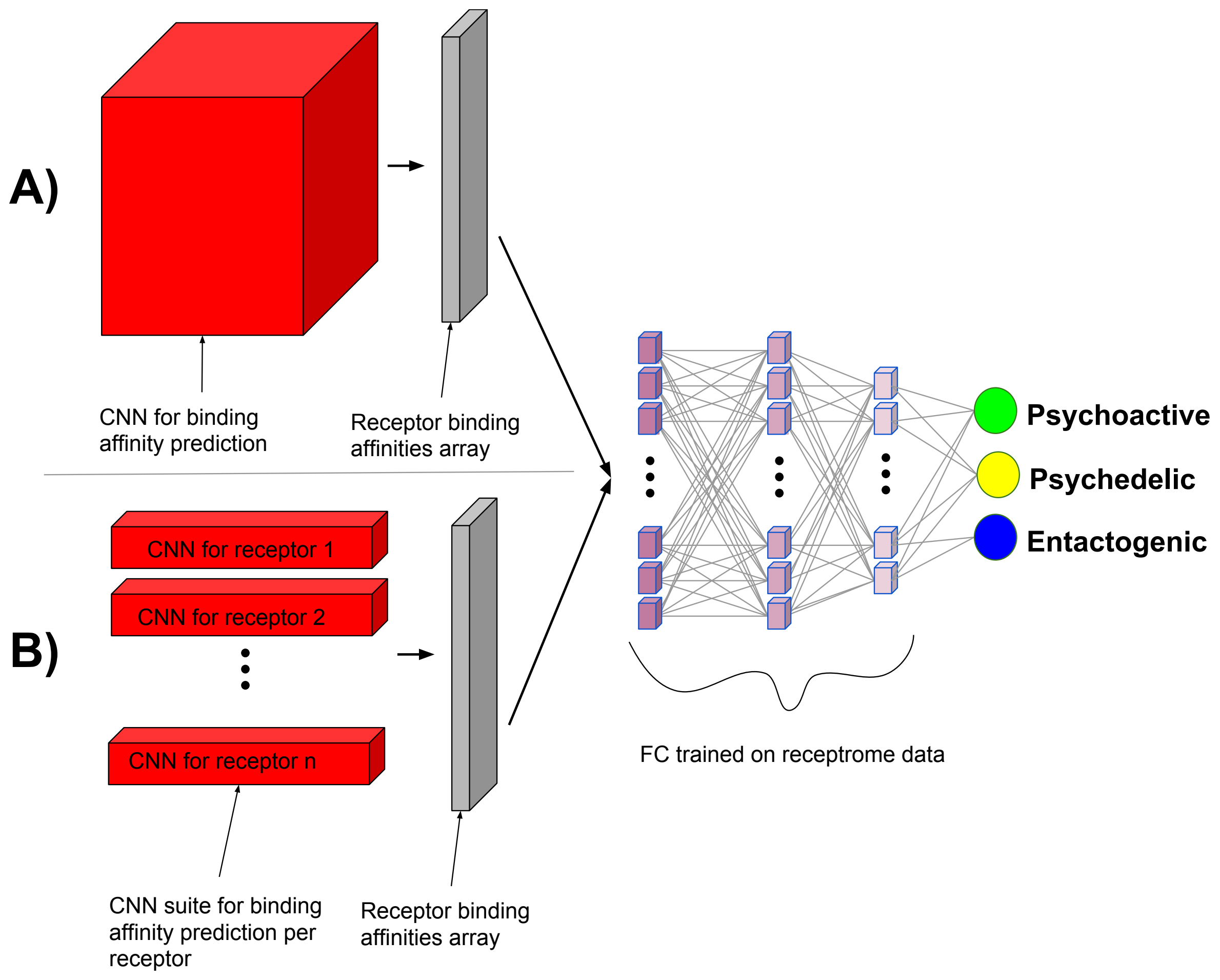
A)
CNN for binding affinity prediction
Receptor binding affinities array
B)
CNN for receptor 1
CNN for receptor 2
CNN for receptor n
CNN suite for binding affinity prediction per receptor
Receptor binding affinities array
Psychoactive
Psychedelic
Entactogenic
FC trained on receptrome data

Fig. 1: Network Architecture

The Data Problem

The largest hurdle encountered on this research was that of data. Deep learning requires a lot of data in order to elucidate and discover the insights and patterns hidden in the data. The problem then, is that such a small handful of ”psychedelic” molecules are known compared to the vast majority which are not psychedelic. The issue is compounded by the fact that there is no clear distinction between ”psychoactive” and ”psychedelic” that is suitable to this task. The easiest distinction is that of psychedelic meaning 5-HT2A receptor agonist but even this falls short as not all such agonists are psychedelic and, as mentioned, Ray found that psychedelic activity has far more to do with the entire receptorome profile than merely with the activation of 5-HT2A alone. Furthermore, because of the unique receptorome profile for each drug it is hard to say when one is psychedelic or not without more analysis into models for Ray's data specifically.

Constructing a data-set then for training becomes the most important next-step for the research. It is important to have a fair ratio of psychedelic to non-psychedelic molecules (say 1:9 or 1:5, but ideally this would also be a hyperparameter) however, determining a metric to use to determine what are the best representative non-psychedelic molecules becomes tricky as they range over every other type and structure of molecule. Furthermore; given that so few psychedelic molecules are known, how many can be afforded to be withheld for cross-validation?

One enticing source of data are the works of Shulgin who created a 1-5 rating scale to describe the subjective dose-dependent effects of different compounds; in his works Pihkal and Tihkal are approximately 500 data points for different phenylethylamines and tryptamines, respectively [6][5]. While this seems a promising source of data, the problem arises in that his rating scale is dose-dependent. The data can be adjusted however, and made useful if it is considered as simply a flag for psychoactivity in a molecule. That being said, despite containing examples of such, the problem of well-representative non-psychedelic data remains however.

Further Research

In addition to the implementation, exploration, and refinement of these methods, once such a suitable model is found it can then be used as a fitness function in identifying ”psychedelic regions” within the chemical space. [4] This is the loftier goal of the research: by studying the chemical space under such a fitness function with suitable dimensions, patterns and groupings of psychedelic molecules can be observed that may lead to better models, distinctions, and predictions as well as an understanding of what makes these compounds so special and different from others.

Remarks

Another issue not discussed thus far is one of computing power. While some of the simpler models (e.g SNN) may not have as hard a time, the large CNNs will require a great deal of computing power to train on. Furthermore, the implementations of Ragoza, et. al's CNN requires GPU optimized machines to compile and execute. Because of this, research on a personal machine is all but impossible, and will require computing resources through WSU or cloud resources such as AWS. The point is moot however, until a suitable dataset can be collected.

Acknowledgements

I would like to acknowledge and thank Dr. Manoranjan and Dr. Nateson for their help and guidance on this project; as well as Washington State University's Department of Mathematics and Statistics for the scholarship and opportunity to investigate this research.

References

[1] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. “DeepDTA: deep drug-target binding affinity prediction”. In: *Bioinformatics (Oxford, England)* (Sept. 2018). URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6129291/`.

[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`.

[3] Thomas S Ray. “Psychedelics and the human receptorome”. In: *PloS one* (Feb. 2010). URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2814854/`.

[4] Jean-Louis Reymond et al. “Chemical space as a source for new drugs”. In: *MedChemComm* (Apr. 2010). URL: `https://pubs.rsc.org/en/content/articlelanding/2010/md/c0md00020e#!divAbstract`.

[5] Alexander T. Shulgin and Ann Shulgin. *Tihkal: the continuation*. Transform Press, 2016.

[6] Alexander Shulgin and Ann Shulgin. *PIHKAL: Phenethylamines I have known and loved: a chemical love story*. Transform Press, 1995.

[7] Kenneth W Tupper et al. “Psychedelic medicine: a re-emerging therapeutic paradigm”. In: *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne* (Oct. 2015). URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4592297/`.