An approach for predicting the effects of endocrine disrupting chemicals on human health using deep learning

Pantelis Karatzas[1], Yiannis Kiouvrekis[2,3], Petros S. Stefaneas[1], Haralambos Sarimveis[1]
[1] National Technical University of Athens

[2] Univrsity of Nicosia

[3] University of Thessaly

Introduction

Recent years have seen considerable development in the fields of machine learning and neural networks in particular [1]. New architectures and methodologies allow for the use of data, such as sound and image, and for drawing conclusions based on these. Developments involving neural networks enable us to successfully address issues such as recognition, image categorization and recognition of objects depicted – in certain instances more efficiently than a human agent could [2].

There is ongoing development in neural network architectures [3]. Increased access to data combined with improved computing power have enabled researchers and programmers to find new applications for neural networks at a very rapid rate [4]. For many years, there have been various attempts at devising systems that can depict chemical structures; 'SMILES' is one of these systems of depiction which is commonly accepted in scientific cirles [5]. The Simplified Molecular-Input Line-Entry System (SMILES) is a standard which takes the form of a linear string describing chemical structures using short ASCII strings. Most molecular processors can be input with SMILES strings and transform those back to 2D depictions or 3D molecular patterns [6, 7] .

Chemical structure images will be used as input in neural networks which will attempt to model their response to human tissue and cells. Our aim was to determine whether this specific information could be used for modelling purposes.

Endocrine Disruptors

The endocrine system plays a central role in regulating metabolism, development, reproduction and behavior in all vertebrates. The hypothesis advanced concerning the presence of endocrine disruptors [8] has led to new studies expressing concerns about the effects of endocrine disruption on health and the environment [9]. Studies incorporate findings and methodologies from different fields, including toxicology, endocrinology, developmental biology, molecular biology, ecology, behavioral biology and epidemiology [9]. An endocrine disruptor is defined as “an exogenous chemical substance or mixture that alters the structure or function(s) of the endocrine system and causes adverse effects at the level of the organism, its progeny, populations, or subpopulations of organisms, based on scientific principles, data, weight-of-evidence, and the precautionary principle” [10]. Data collected from ecological studies, animal models, clinical observation of human subjects and epidemiological studies indicate that endocrine disrupting chemicals pose a significant risk to wild life and human health [11].

Neural Networks

In recent years, neural networks, especially those exhibiting synergistic properties, have been at the cutting edge of image processing, producing very good results. So far, they have been able to successfully address issues of classification, recognition and recognition of objects depicted on images. Therefore, chemical structure images will be used as an input in neural networks with the end goal of classifying them according to their classes.

Neural Network Architectures

LeNet and AlexNet

In terms of contemporary standards, LeNet-5 is a very simple network. LeNet-5 exhibits many interesting architectural choices, not too common in today’s age of deep learning. The AlexNet architecture consists in Convolution, max-pooling, Local Response Normalization (LRN) and fully connected (FC) layer [12]. It is very similar to the LeNet network except for the fact that it contains more layers in total. The main difference is the

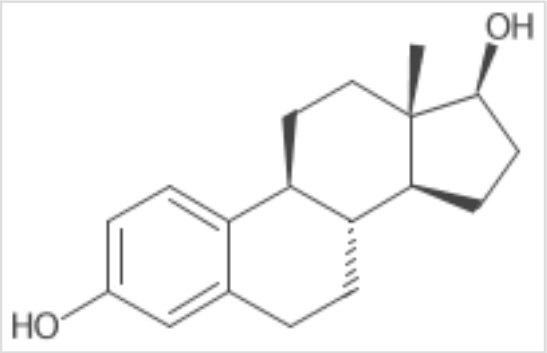
activating function employed, namely the ReLU function. AlexNet comprises eight layers; five of these are convolutional layers, some of which are followed by max-pooling layers, and three are fully connected layers. The ReLU activation function utilized by AlexNet shows improved training performance compared to tanh or sigmoid.

Data Set

The data set consists of 1,459 chemical structures. Based on experimental data, they have been labeled with values concerning their Relative Binding Affinity on a logarithmized scale (logRBA). The data were gathered from the EADB dataset (Estrogenic Activity Database) [15, 16, 17]. The data subset used involves only the endpoints for species (human) and for logRBA.

Deep neural networks are very effective in dealing with image classification problems. In order to proceed with our modelling, we classify the data in 3 classes according to their experimental response. The first class comprises structures which have response values in the [-3.328, -0.26] range, the second class comprises response values in the [-0.259, 0.824] range and the third comprises values in the [0.826, 2.857] range. The classes have been encoded as a One-hot vector.

Afterwards, the images of the chemicals were generated using the chemistry development kit (CDK) [13]. The image generating software plays a significant role. The first experiments were conducted with the usage of two kinds of images The experiment utilized CDK generated images. In order to increase data size we supplied additional indigo generated images. However, this did not contribute to neural network learning but, on the contrary, it prevented convergence. [14]. The reason for generating two kinds of images was dataset augmentation, since the starting dataset is relatively small for deep neural network training. In the end we used only the CDK generated images since we could not train the model with both image types.



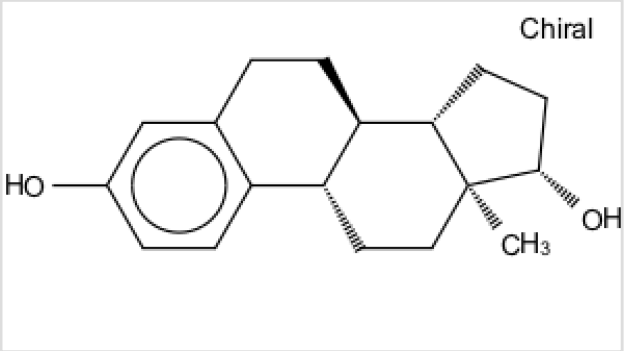


Figure 1: Two smiles images generated using two different softwares. Left: CDK produced SMILE image; Right: Indigo produced SMILES image.

The images have been resized and pasted into white background so as to fit into squares as the input to a neural network. The dimensions used in modelling are 128 * 128, 200 * 200 and 256 * 256. 128 and 256 sized pictures favour the pooling layers of a deep neural network that can downsample the images to 8 * 8 sized convolutional kernels, or even 4*4. This way we can add many layers on the neural network if needed.

In order to proceed with training, data is split into random batches according to training needs. The memory capacity of the equipment available to the user plays a crucial role. In our case we conducted random sampling from batches of 42 images each. The batches are fed into the neural network to proceed with the training procedure.

It should be noted that images have been normalized. Their pixel values were scaled so as to have a mean value of 0 and a standard Deviation of 1.

Results and discussion

In this section we present the results of our approach. We proceeded employing two architectures of neural networks. In one instance we used architectures of the ImageNet

type, as previously described, and in the second instance we used neural networks with Residual blocks.

Modelling through ImageNet

The data presented consist of images of chemical molecules in 2D format. After taking into account the image size, we used the respective neural network. The dimensions of the images as mentioned were 128 * 128 or 200 * 200 and 256 * 256. The neural network was developed in a similar manner. Three convolutional layers were following the network input. Each convolutional layer is followed by a 2 * 2 pooling layer created by downsampling of the network input. Therefore, depending on the input we have the final convolutional layer with 16 * 16 filters times the filter depth (e.g. 56), 25 * 25 times the filter depth and, in the case of the 256 *256 input, 32 * 32 times the filter depth. The layering procedure follows together with the two final fully connected network layers which consist of 1024 nodes according to our modelling. At the network output we have the three classes we need to predict. The activation function employed in all neural networks is ReLU (Rectified Linear Unit)

$$f(x) = \max(0, x)$$

We used two optimization functions, the classic Gradient Descent method and AdamOptimizer. The learning rate value was relatively high and constant, namely 0.3. On every network layer we used the dropout method to avoid overfitting the model.

It was noted that the learning rate value played a special role in training. The lower this value is, the more difficult it becomes for the model to converge, while in order for training to begin and for convergence to proceed, the training procedure will have to commence at a relatively high value – at least 0.3. The input image size was not significant, since the accuracy of the model was affected by 1% at most, without exhibiting any patterns.

The training accuracy graphs are presented below. The first graph (Figure 1) shows the development of accuracy during data training. These are the same data with which the models are trained. As anticipated, from a certain point onwards (i.e. after step 5000) accuracy approaches 1:

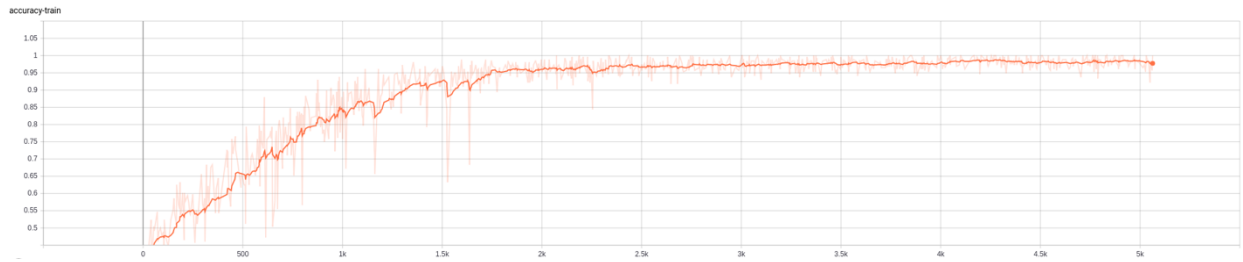


Figure 2

The following (Figure 2) is an accuracy graph of a validation data model. In this instance we note that accuracy reaches 70%. The model used was an ImageNet type model and resulted in one of the best model performances we recorded during training.

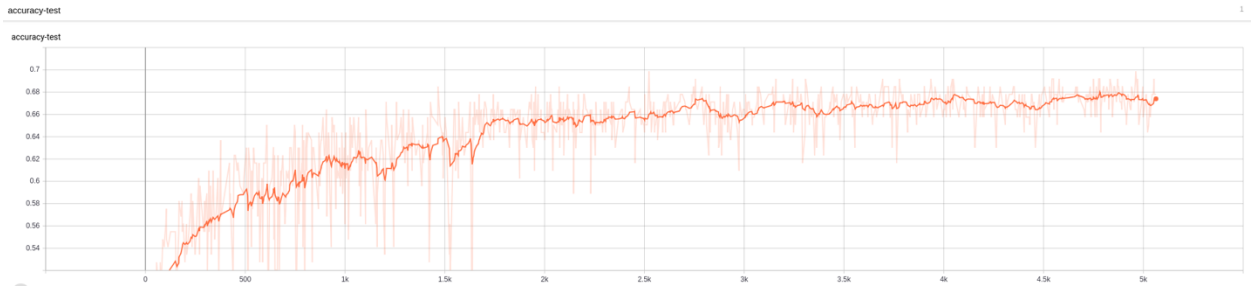


Figure 3

In the field of machine learning and, in particular, concerning the issue of statistical classification, a confusion matrix, also known as an “error matrix” [7], refers to a certain matrix configuration which allows for visualizing an algorithm’s performance, typically a case of supervised learning. Each row in the matrix represents occurrences in a predicted class, while each column represents occurrences in an actual class (or vice versa) [2]. It is named after the fact that it facilitates determining whether the system causes confusion in two categories (i.e. typically a wrong flagging between the two).

The following table is a confusion matrix of predictions with the accuracy at 0.67%.

n=146	Number of images	Predicted class 0	Predicted class 1	Predicted class 2

n=146	Number of images	Predicted class 0	Predicted class 1	Predicted class 2
Actual class 0	53	**34 (0.64%)**	13	6
Actual class 1	39	6	**26 (0.66%)**	7
Actual class 2	54	4	12	**38 (0.70)**

Figure 4

Matthews Correlation Coefficient (MCC), introduced by biochemist Brian W. Matthews in 1975, is used in machine learning to measure the quality of binary classifications (two categories). Although MCC is equivalent to Karl Pearson's phi-coefficient, introduced decades after, it is still widely used in bioinformatics.

The coefficient takes into account both true and false positives and negatives and is generally considered as a balanced measure which can be employed even if the classes differ greatly in size. In essence, MCC is a correlation coefficient applied to observed and predicted binary classifications. It yields a value between -1 and 1. A coefficient of 1 represents a perfect perdiction, 0 is better than a random prediction and -1 indicates complete disagreement between prediction and observation. Our value for the model's predictions is MCC = 0.51 .

MCC = 0.51

Modelling with Residual nets

Residual nets are an alternative to the neural networks architecture. In our study we attempted to use them, since they show very good results in a number of modelling cases. We tried various levels in the network. The training accuracy graphs are presented below (Figure 5).

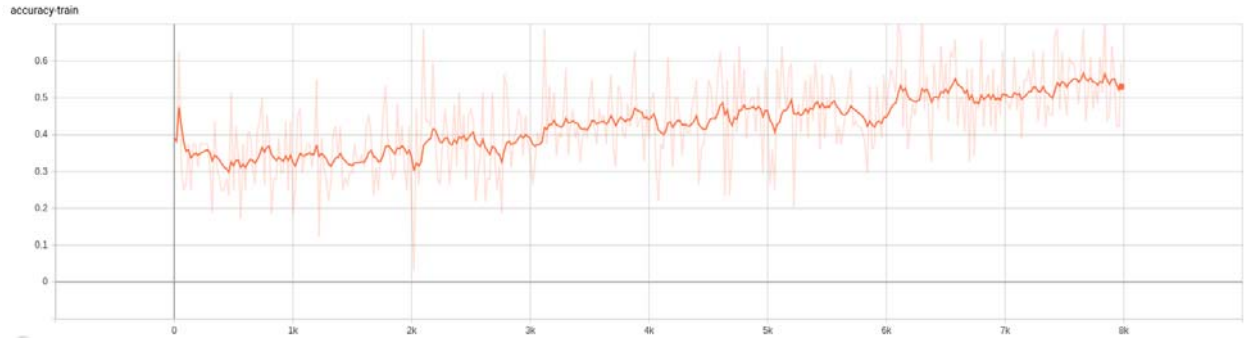


Figure 5

The following graph (Figure 6) is an accuracy graph of a validation data model.

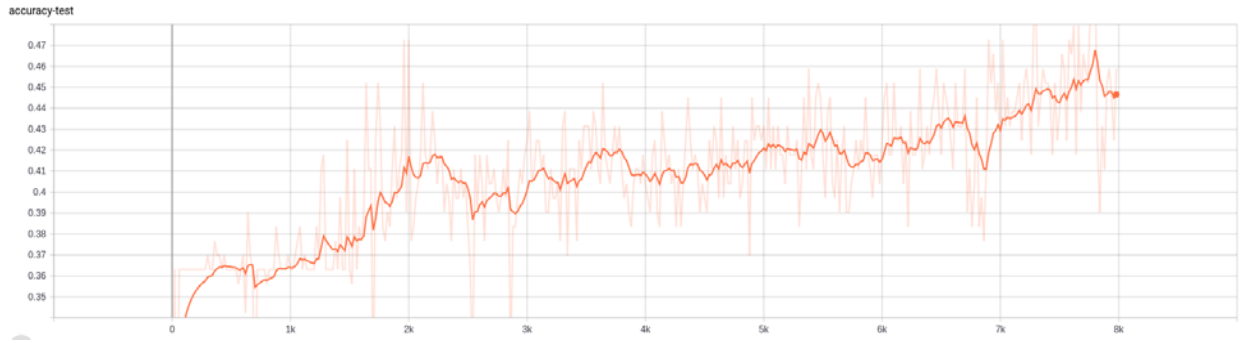


Figure 6

Based on the graphs, we note that training for these particular models was not complete, but, given the greater multitude of parameters, training time tends to be a prohibiting factor. It took approximately 6 hours to reach step 8000 (as shown in the following) on an 8gb memory capacity card. Furthermore, the accuracy levels recorded do not make us optimistic about attaining much better results from ImageNet neural networks. We will need a more powerful computer and access to more resources to confirm that claim.

Further Discussion and Research

In this paper we showed that 2D structures of chemical compounds can be a significant piece of information in QSAR modelling. Furthermore, we succeeded in finding certain network architectures capable of using this piece of information in the best possible manner.

Subsequently, we could proceed with studying which other responses could be modeled following this particular process as well as with collecting more extensive data, since more data lead to better results in neural network modelling.

We also understand that a neural network's input can take any shape we need and that in addition to a molecule's 2D structure there is a 3D structure, which provides more detailed information, such as indicating a particular atom's position relative to that of other atoms withing the structure. Suitable encoding which will allow for the use of data formation by neural networks could possibly also lead to better results. Furthermore, the creation of larger data sets will provide models with better predictive capacity.

[1] Oludare Isaac Abiodun, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, Humaira Arshad, State-of-the-art in artificial neural network applications: A survey, Heliyon, Volume 4, Issue 11, 2018, e00938, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2018.e00938.

[2] Andina, Diego and Voulodimos, Athanasios and Doulamis, Nikolaos and Doulamis, Anastasios and Protopapadakis, Eftychios, *Deep Learning for Computer Vision: A Brief Review* , Computational Intelligence and Neuroscience

[3] Wang, X., Zhao, Y. & Pourpanah, F. Recent advances in deep learning. *Int. J. Mach. Learn. & Cyber.* **11,** 747–750 (2020). https://doi.org/10.1007/s13042-020-01096-5

[4] https://openai.com/blog/ai-and-compute/

[5] Hessler G, Baringhaus KH. Artificial Intelligence in Drug Design. *Molecules*. 2018;23(10):2520. Published 2018 Oct 2. doi:10.3390/molecules23102520

[6] Weininger, D., Weininger, A., Weininger, J.L.: SMILES 2: Algorithm for Generation of Unique SMILES Notation, Daylight Chemical Information Systems, Irvine, California 92714 (1989); Note that although the Unique SMILES implementation has been changed by the Daylight Chemical Information System, this appears to be the most recent publication describing the algorithm

[7] Weininger, D.: SMILES, a Chemical Language and Information System 1: Introduction to Methodology and Encoding Rules, Medicinal Chemistry Project, Pomona College (1988)

[8] Colborn, T.; vom Saal, F.S.; Soto, A.M. Developmental effects of endocrine-disrupting chemicals in wildlife and humans. Environ. Health Perspect. 1993, 101, 378–384.

[9] Myers, J.P.; Guillette, L.J., Jr.; Palanza, P.; Parmigiani, S.; Swan, S.H.; vom Saal, F.S. The emerging science of endocrine disruption. In International Seminars on Planetary Emergencies, 30th Session; Ragaini, R.C., Ed.; World Scientific Publishing: London, UK, 2004; pp. 105–121

[10] Zoeller, R.T.; Brown, T.R.; Doan, L.L.; Gore, A.C.; Skakkebaek, N.E.; Soto, A.M.; Woodruff, T.J.; Vom Saal, F.S. Endocrine-disrupting chemicals and public health protection: A statement of principles from The Endocrine Society. Endocrinology 2012, 153, 4097–4110.

[11] Street, M.E.; Angelini, S.; Bernasconi, S.; Burgio, E.; Cassio, A.; Catellani, C.; Cirillo, F.; Deodati, A.; Fabbrizi, E.; Fanos, V.; Gargano, G.; Grossi, E.; Iughetti, L.; Lazzeroni, P.; Mantovani, A.; Migliore, L.; Palanza, P.; Panzica, G.; Papini, A.M.; Parmigiani, S.; Predieri, B.; Sartori, C.; Tridenti, G.; Amarri, S. Current Knowledge on Endocrine Disrupting Chemicals (EDCs) from Animal Biology to Humans, from Pregnancy to Adulthood: Highlights from a National Italian Meeting. *Int. J. Mol. Sci.* **2018**, *19*, 1647.

[12] Alom, M.Z.; Taha, T.M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M.S.; Hasan, M.; Van Essen, B.C.; Awwal, A.A.S.; Asari, V.K. A State-of-the-Art Survey on Deep Learning Theory and rchitectures. *Electronics* **2019**, *8*, 292.

[13] Willighagen et al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. J. Cheminform. 2017; 9(3), doi:10.1186/s13321-017-0220-4

[14] https://lifescience.opensource.epam.com/download/indigo/index.html

[15] Jie Shen, Lei Xu, **Hong Fang**, Ann M Richard, Jeffrey D Bray, Richard S Judson, Guangxu Zhou,Thomas J Colatsky, Jason L Aungst, Christina Teng, Stephen C Harris, Weigong Ge, Susie Y Dai, Zhenqiang Su, Abigail C Jacobs, Wafa Harrouk, Roger Perkins, Weida Tong, Huixiao Hong. "EADB: An Estrogenic Activity Database for Assessing Potential Endocrine Activity," *Toxicological Science 2013,* 135(2), 277-291. (PMID: 23897986 Abstract).

[16] Hui Wen Ng, Roger Perkins, Weida Tong, and Huixiao Hong. "Versatility or Promiscuity: The Estrogen Receptors, Control of Ligand Selectivity and an Update on Subtype Selective Ligands." *International Journal of Environmental Research and Public Health* 2014, 11(9):8709-8742. (PMID: 25162709 Abstract).

[17] Hui Wen Ng, Wenqian Zhang, Mao Shu, Heng Luo, Weigong Ge, Roger Perkins, Weida Tong, Huixiao Hong. "Competitive molecular docking model for predicting estrogen receptor agonists and antaonists." *BMC Bioinformatics* 2014, 14(11), S4.