OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Utilizing QSAR Prediction Model to Identify Potential HER2 Inhibitors

Dung Pham Linh and Thu Dang Hanh Minh
sallypham281@gmail.com and danghanhminhthu08@gmail.com

ABSTRACT

Context: The HER2 receptor is a transmembrane protein that regulates the growth and division of cells. With other EGFR (epidermal growth factor receptor) family members, it forms homodimers or heterodimers in a ligand-dependent and -independent manner, activating signaling pathways that trigger cell proliferation, survival, and migration[1]. In some cases, if there are too many copies of the receptor, the cells will grow and divide uncontrollably and leading to more aggressive forms of cancer cells[2]. This paper anticipates the potential antagonist compounds via the Google Colab platform to evaluate their potential and suggest some HER2 inhibitors.

Objective: The research performs a ligand-based virtual screen experiment using the Artificial Neural Network-based Quantitative Structure-Activity Relationship (ANN-QSAR) to screen a large library of 50,000 molecules to identify potential HER2 inhibitors.

Conclusion: There are 2305 inactive compounds in the dataset and 1760 active compounds extracted from the CHEMBL database. These compounds have a smaller and more flexible structure when compared with other small molecule inhibitors like the (TKIs) tucatinib, lapatinib, and neratinib, which seem to have a larger and more rigid ring system [3]. These inhibitors are important to the treatment of breast cancer. Yet, Herceptin is still needed because it is more well-known as the single most important treatment for breast cancer in both metastatic and neoadjuvant settings [4].

INTRODUCTION

The gene mutations you're born with and those that you acquire throughout your life build how your cells function every day. Yet sometimes, they can overexpress and cause cancer. The HER2 is a receptor tyrosine-protein kinase erbB-2 protein that belongs to the epidermal growth factor receptor (HER) family of receptors [5]. In the development of some cancers, they regulate cell growth, survival, and differentiation through multiple signal transduction pathways [6]. Human epidermal growth factor receptor 2, 1 of 4 membrane tyrosine kinases (TKs), was found to be amplified in a human breast cancer line 25 years ago, and this amplification played an important role in the pathogenesis and development of breast cancer 2 years later [7].

The use of studying inhibitors that work inside the cells is vital to our understanding and prevention of cancer cells in the future. Some of the smaller molecule HER2 inhibitors that are FDA-approved includesTKIs neratinib, followed by tucatinib, and then lapatinib. Neratinib shows the highest potent of these. It's an inhibitor of HER2, EGFR, and HER4, where it binds covalently to interfere with cancer cells' division, growth, survival, and invasion. The key functional group of neratinib is an α,β-unsaturated carbonyl (an acrylamide moiety) that acts as a reactive "warhead." This group forms a covalent bond with a cysteine residue in the HER family receptors to make neratinib permanently attached to the enzyme's active site[8]. This explains why the drug can shut off cells even when it's not present in plasma [9]. Clinically, the utilization of neratinib is beneficial and effective, used for both early and advanced stages, but its strong action can lead to some serious side effects that can harm patients' kidneys, liver, and cause severe diarrhea [10].

Apart from neratinib's popularity, tucatinib is a new potential for an effective drug [11]. Contrary to neratinib, it is highly selective for advanced and metastatic HER2+ breast cancer and collateral cancer [12]. Used in conjunction with medicine trastuzumab (Herceptin) and capecitabine (Xeloda), the drug will target breast cancer that has already spread to other parts of the body and aren't treatable via surgery for adults with at least 1 chemotherapy treatment. Tucatinib is a non-covalent inhibitor, so it doesn't have the electrophilic acrylamide "Michael acceptor" present in pan-HER inhibitors like neratinib. While neratinib can irreversibly alkylate a cysteine in the HER family kinase domain, tucatinib doesn't have covalent-binding warhead and reversibly binds HER2, causing it to be highly potent against HER2 while sparing EGFR to minimize EGFR-related toxicities (such as severe rash) seen with less selective inhibitors[13]. Its results will help patients live longer and block the activity of cancer [14]. Still, some of side effects might include vomiting, nausea, or weight loss overall[13].

Lapatinib is an older TKI used in combination with the chemotherapy drug capecitabine to target advanced or metastatic cancer (cancer that already spread) in patients who have already received other treatments (eg,anthracycline,taxane, trastuzumab) that don't work well [15]. It binds to the intracellular phosphorylation to prevent ligand binding from receptor autophosphorylation; [16] and has 4 key functional include a 4-anilinoquinazoline scaffold, a flat ring system that anchors the drug in the kinase pocket, and several substituents that enhance its activity. Notably, lapatinib contains halogen atoms (chlorine and fluorine) on its aromatic rings to improve binding, as well as a methylsulfonyl-ethylamino moiety on its structure. The methylsulfonyl group is a polar functional unit that increases lapatinib's water solubility and helps the drug extend into the solvent-exposed region of the binding site [17]. These structural features give lapatinib strong dual activity against HER2 and EGFR, albeit missing neratinib permanent covalent bond. While Lapatinib's more transient binding generally leads to tolerable toxicity in patients, serious cases might include liver disease or dry and flaky skin.[16]There are still many unanswered questions about their toxicity and efficacy.

On the surface level, Herceptin in the treatment has been demonstrated in multiple clinical studies. Herceptin is not a small molecule at all, but a monoclonal antibody, designed to seek out and bind HER2 receptors on the surface of cancer cells. Instead of a traditional chemical functional group, Herceptin's

antigen-binding fragment (Fab) region is specifically used to recognize and attach to a portion of HER2 protein, blocking the cancer cell's ability to grow. In clinical studies, Herceptin therapy can significantly improve survival rates and reduce the risk of recurrence for patients with HER2-positive breast cancer.[18] Nevertheless, this method has its own risks like reduced heart function or even heart failure.[19]

To predict the potential antagonist compounds for HER2, we will use a virtual screen experiment to identify the efficacy of compounds and suggest several effective ones.

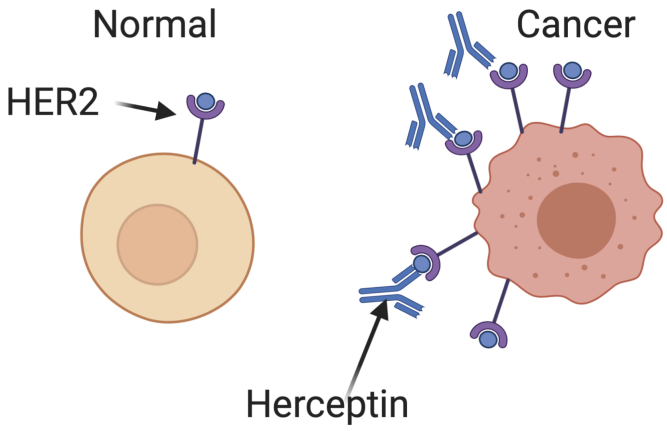

Figure 1. Mechanism of action of Herceptin

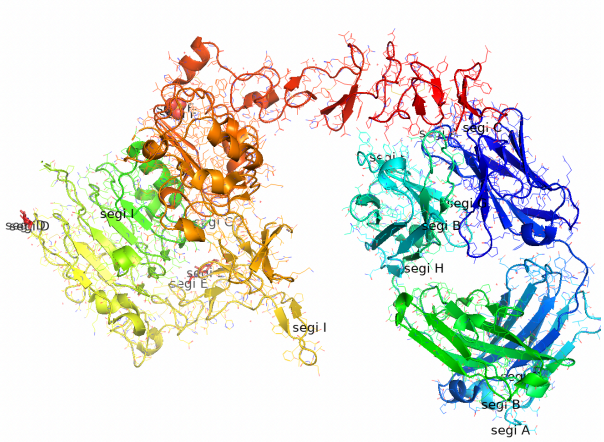

Figure 2. Crystal structure of extracellular domain of human HER2 complexed with Herceptin Fab

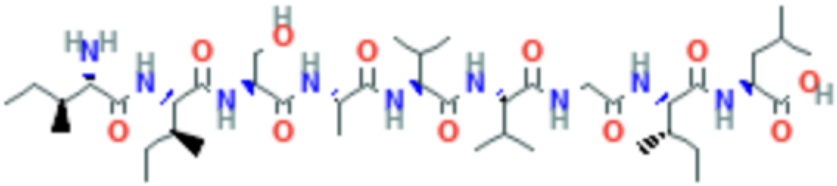

Figure 3. HER-2 structure

METHODS

The research was carried out mostly on Google Colab (https://colab.research.google.com/) using Python 3.12, relying on several Key Python libraries, including NumPy for numerical operations, Pandas for data preprocessing, Scikit-learn for data splitting and evaluation, and Matplotlib for visualization. Building on this setup, the artificial neural network (ANN) model was then implemented and trained using TensorFlow/Keras.

Firstly, we used the RDKit library, which is widely used for cheminformatics tasks, to extract the compounds for this study.

The extraction process involved these steps:

- Retrieved the ChEMBL ID of HER2 (CHEMBL1824) from the ChEMBL database. (https://www.ebi.ac.uk/chembl)
- Extracted molecules tested against HER2.
- Removed compounds with invalid or unusual molecular structures.
- Eliminated duplicate compounds to avoid bias.
- Obtained a curated dataset suitable for QSAR modeling for the result.

Before structure filtering, we include a criterion that only includes compounds with reported biological activity (e.g., IC50, Ki) within a relevant range.

Next, we use the Chem functions on the Google Colab platform and the RDKit 2023.09.1 package (http://www.rdkit.org) to analyze our compounds. The relevant compounds are categorized into inactive compounds and active compounds, an evaluation based on the value of pKi. While active compounds have a pKi value greater than or equal to 6 and less than 15, the rest of the compounds with pKi values not in this range are inactive compounds (pKi values smaller than 6). From this processed dataset, we utilize Morgan circular fingerprints (ECFP4) with a radius of 2 generated as input for the QSAR prediction model [20].

Building on these fingerprints, the ANN-QSAR model uses a predictive model for the chemical structure and the biological activity of approximately 50,000 compounds sourced from the eMolecules database. The model evaluates from 1024 input features to the first hidden layer of 32 neurons with ReLU activation function, followed by a 1-neuron output layer with Sigmoid activation function at the end. In this paper, we implemented a small dropout rate of 0.5 to provide light regularization while still allowing the model to learn effectively [21].

With the data set containing the molecular structures and biological activities of HER2 prepared, it is then randomly split into a training set and a test set with the ratio 80/20. This provides a more realistic assessment of the model's generalizability to new chemical structures. The classification model is then trained using the Keras package (version 3.10.0) and outputs the data results in accuracy, precision, and recall of the ANN models on both the training and the test sets.

Finally, virtual screening is used to screen a large library of 50,000 molecules from eMolecules database to identify potential HER2 inhibitors.

RESULTS

We have extracted 5,169 compounds from the ChEMBL database to obtain molecules with reported biological activity against HER2, ensuring relevance for QSAR modeling. After filtering and classification, we also received - 2,529 compounds labeled as active ($pKi \geq 6$) and 2,640 as inactive ($pKi < 6$).

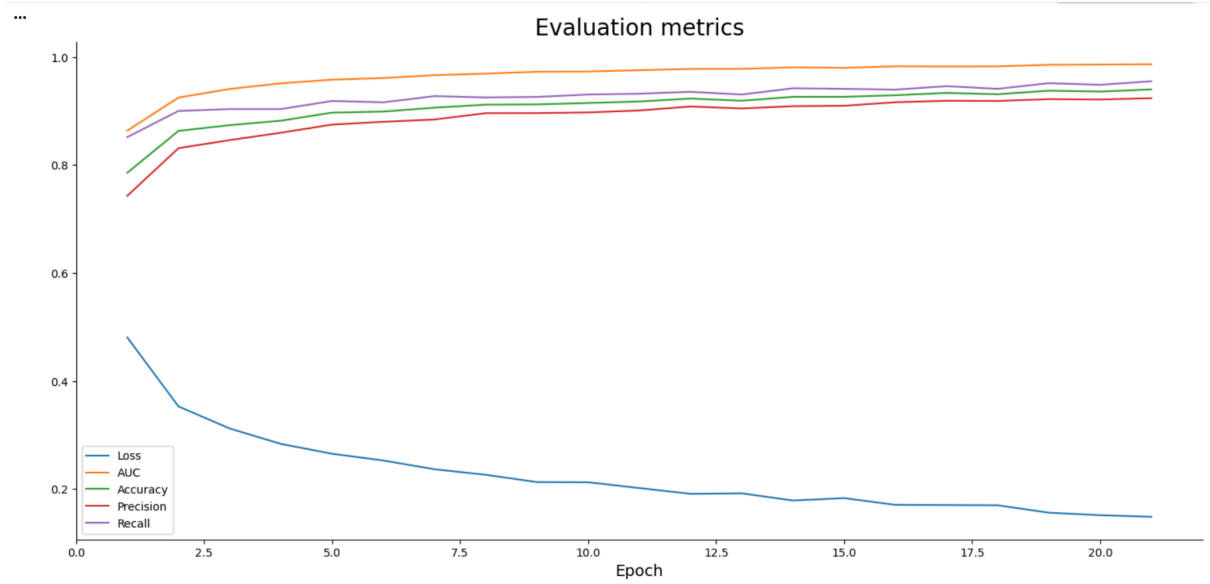

Figure 4. Evaluation metrics in training QSAR model

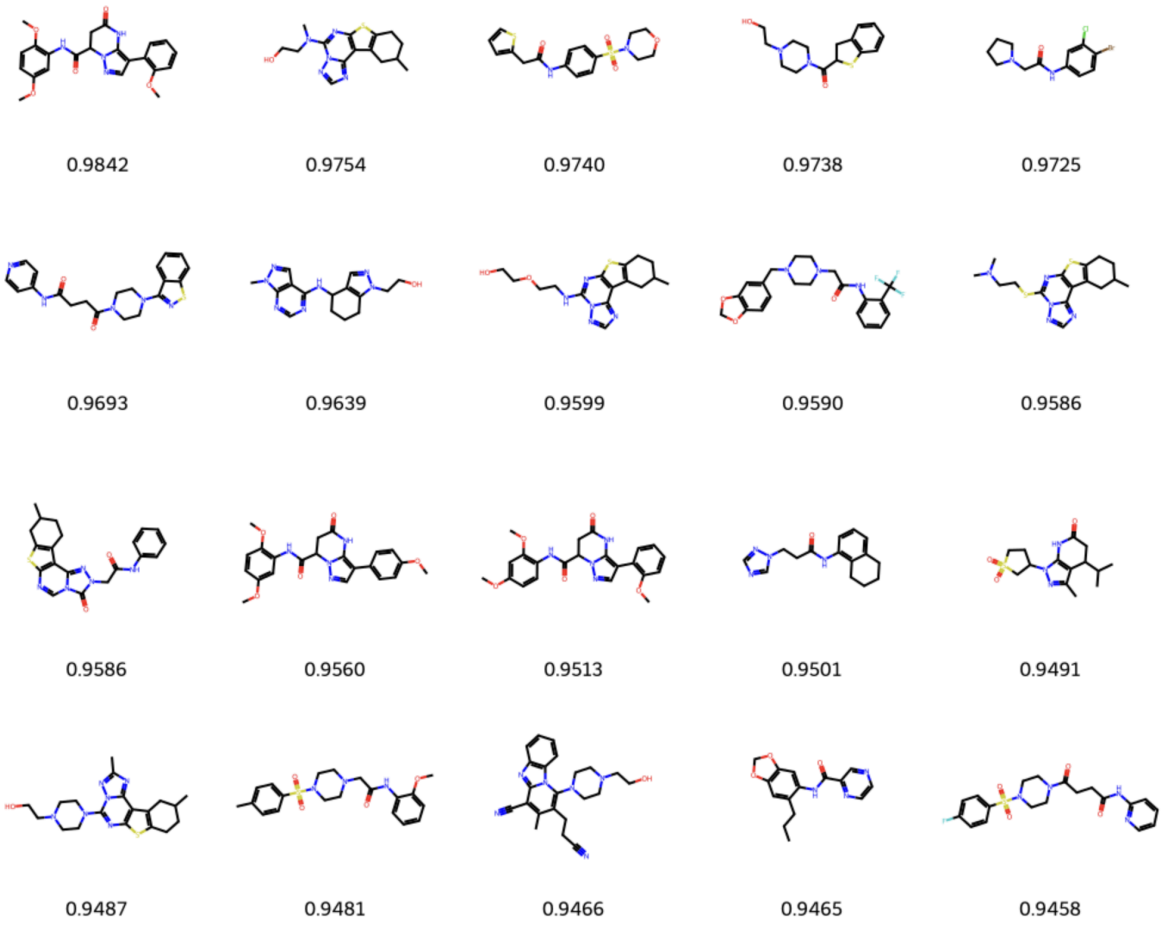

Figure 5. Top 20 compounds with high predicted value (higher means more likely to be active) in biological activity related to HER2

Under each molecular structure are the numbers indicating the model's confidence in predicting specific properties, such as bioactivity or binding. A score approaching 1.0 suggests that the molecule satisfies the target criteria, while a score near 0 proves otherwise.

DISCUSSION

During the screening, ANN performed an excellent overall performance by reaching close-to-perfect output for the molecules inside the dataset. In other words, it efficiently captures consistently high values,

mostly above 0,9… displaying the relationship among molecular fingerprints and biological activity. As a result, the network architecture, together with the hidden layers, activation functions, and possibly regularization techniques, fits the dataset properly, minimizes underfitting, and suggests solid predictions.

As for the top-ranked compounds identified from the ANN screening percentage, we notice numerous common functional groups and structural motifs like heterocyclic rings containing nitrogen and sulfur, sulfonamide moieties, and carbonyl groups. Many among them are combined with bicyclic or polycyclic scaffolds, which are typical in kinase inhibitors. Comparing them to recognised HER2 inhibitors, these molecules maintain core pharmacophoric functions that facilitate binding to the kinase domain, a characteristic of hydrogen bond donors/acceptors and hydrophobic aromatic rings. However, some top applicants display specific substitutions and ring structures not found in approved drugs, suggesting they might provide novel interactions and possible improvement in selectivity or potency.

In short, the screening outcomes both highlight a convergence on key functional motifs essential for HER2 inhibition and introduce structural range.

LIMITATIONS

While our model has successfully identified promising candidates for HER2 treatment, scientific integrity requires us to address the inherent challenges in this computational approach.

Firstly, we must acknowledge the "False Positive" paradox. Our model was trained using molecular fingerprints, which are essentially 2D representations of chemical structures. However, biological interactions occur in a complex, 3D space. There is a risk that the AI is identifying compounds that appear effective on a 2D plane but may physically fail to fit the HER2 binding pocket in a living system.

Secondly, the reliability of a model depends on the quality of the data it uses for learning. By utilizing the ChEMBL database, we compiled information from various laboratories. Differences in experimental conditions and research protocols across laboratories can also introduce noise and lead to inconsistent results.

Moreover, while the model’s strong performance on the test dataset is encouraging, it is not sufficient to conclude about real-world applications. To truly evaluate their effectiveness, these compounds must be tested directly on actual cancer cell lines.

CONCLUSION

In this study, deep learning helped save both time and resources by filtering out thousands of unsuitable candidates at an early stage. However, it should be viewed as a supportive tool rather than a replacement for laboratory experimentation. Overall, the findings offer a strong starting point for further 3D analysis and, eventually, clinical trials.

References:

[1] Iqbal, Nida, and Naveed Iqbal. "Human Epidermal Growth Factor Receptor 2 (HER2) in Cancers: Overexpression and Therapeutic Implications." Molecular Biology International 2014, no. 1 (September 7, 2014): 1-9

[2] Gutierrez, Carolina, and Rachel Schiff. "HER2: Biology, Detection, and Clinical Implications." Archives of Pathology & Laboratory Medicine 135, no. 1 (January 1, 2011): 55–62. https://doi.org/10.5858/2010-0454-rar.1.

[3] Yarden, Yosef. "Biology of HER2 and Its Importance in Breast Cancer." Oncology 61, no. 2 (2001): 1–13. https://doi.org/10.1159/000055396.

[4] BREASTCANCER.ORG. "Herceptin." www.breastcancer.org, 2024. https://www.breastcancer.org/treatment/targeted-therapy/herceptin

[5] Cheng, Xiaoqing. "A Comprehensive Review of HER2 in Cancer Biology and Therapeutics." Genes 15, no. 7 (July 11, 2024): 903–3. https://doi.org/10.3390/genes15070903.

[6] Conlon, Neil T., Jeffrey J. Kooijman, Suzanne J. C. van Gerwen, Winfried R. Mulder, Guido J. R. Zaman, Irmina Diala, Lisa D. Eli, Alshad S. Lalani, John Crown, and Denis M. Collins. "Comparative Analysis of Drug Response and Gene Profiling of HER2-Targeted Tyrosine Kinase Inhibitors." British Journal of Cancer 124, no. 7 (March 1, 2021): 1249–59. https://doi.org/10.1038/s41416-020-01257-x.

[7] Subramanian, Ashok, and Kefah Mokbel. "The Role of Herceptin in Early Breast Cancer." International Seminars in Surgical Oncology 5, no. 1 (April 28, 2008). https://doi.org/10.1186/1477-7800-5-9.

[8] [19-8']Karim Aljakouch, Tatjana Lechtonen, Hesham K Yosef, Mohamad K Hammoud, Wissam Alsaidi, Carsten Kötting, Carolin Mügge, Robert Kourist, Samir F El‐Mashtoly, and Klaus Gerwert. 2018. "Raman Microspectroscopic Evidence for the Metabolism of a Tyrosine Kinase Inhibitor, Neratinib, in Cancer Cells." Angewandte Chemie 57 (24): 7250–54. https://doi.org/10.1002/anie.201803394.

[9] Schlam, Ilana, and Sandra M. Swain. "HER2-Positive Breast Cancer and Tyrosine Kinase Inhibitors: The Time Is Now." Npj Breast Cancer 7, no. 1 (May 20, 2021). https://doi.org/10.1038/s41523-021-00265-1

[10] https://www.facebook.com/Drugscom "Neratinib Uses, Side Effects & Warnings." Drugs.com, 2024. https://www.drugs.com/mtm/neratinib.html

[11] Musechem.com. “Tukysa (Tucatinib): A Novel Treatment for HER2-Positive Breast Cancer,” 2024.
https://www.musechem.com/blog/tukysa-tucatinib-a-novel-treatment-for-her2-positive-breast-cancer/
[12] Stark, Roderick T, Dominic R Pye, Wenyi Chen, Oliver J Newton, Benjamin J Deadman, Philip W Miller, Jenny-Lee Panayides, Darren L Ril ey, Klaus
[13] Gentile, Gabriella, Simone Scagnoli, Luca Arecco, Daniele Santini, Andrea Botticelli, and Matteo Lambertini. 2024. “Assessing Risks and Knowledge Gaps on the Impact of Systemic Therapies in Early Breast Cancer on Female Fertility: A Systematic Review of the Literature.” Cancer Treatment Reviews 128 (July): 102769. https://doi.org/10.1016/j.ctrv.2024.102769.
[14] Hellgardt, and King Kuok. “Assessing a Sustainable Manufacturing Route to Lapatinib.” Reaction Chemistry & Engineering 7, no. 11 (January 1, 2022): 2420–26.
https://doi.org/10.1039/d2re00267a
[15] Medlineplus.gov. “Tucatinib: MedlinePlus Drug Information,” 2020.
https://medlineplus.gov/druginfo/meds/a620032.html
[16] https://www.facebook.com/Drugscom. “Lapatinib Advanced Patient Information.” Drugs.com, 2025. https://www.drugs.com/cons/lapatinib.html
[17] “Lapatinib | 231277-92-2.” 2016. ChemicalBook. 2016.
https://www.chemicalbook.com/ChemicalProductProperty_EN_CB8855402.htm.
[18] Greenblatt, Karl, and Karam Khaddour. 2024. “Trastuzumab.” Nih.gov. StatPearls Publishing. June 22, 2024. https://www.ncbi.nlm.nih.gov/sites/books/NBK532246/?utm_source=chatgpt.com.
[19] Bradley, Rosie, Jeremy Braybrooke, Richard Gray, Robert Hills, Zulian Liu, Richard Peto, Lucy Davies, et al. 2021. “Trastuzumab for Early-Stage, HER2-Positive Breast Cancer: A Meta-Analysis of 13 864 Women in Seven Randomised Trials.” The Lancet Oncology 22 (8): 1139–50.
https://doi.org/10.1016/S1470-2045(21)00288-6.
[20] Morgan, H. L. “The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service.” Journal of Chemical Documentation 5, no. 2 (May 1, 1965): 107–13. https://doi.org/10.1021/c160017a018
[21] Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, and Ruslan Salakhutdinov. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting.” Journal of Machine Learning Research 15 (2014): 1929–58.
https://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf?utm_content=buffer79b4