Using active learning to design the optimal lab experiments needed to improve antibody-antigen binding prediction

Dominique Weltevreden[1,2]

Supervised by prof. dr. Antoine van Kampen[1,3], dr. ir. Perry Moerland[3] and Daria Balashova[3]

[1] University of Amsterdam
[2] UMC Utrecht
[3] Amsterdam UMC

Keywords: Active learning · Immunology · Convolutional Neural Networks

1 Abstract

1.1 Introduction

Antibody drug therapy is an important method of treatment in oncology, immunology and haematology [5][12]. Antibodies bind to a specific target antigens, but mutations in these antigens can disrupt this binding [6]. Once an antibody loses effectiveness to a variant of a virus, other antibody cocktails might remain effective to bind to this new antigen variant [11].

However, testing each available antibody against a new antigen rapidly increases costs. Furthermore, it can quickly become unfeasible as multiple mutations might occur at once [11]. In such cases, machine learning models can be used to predict which other antibodies might bind the mutated antigen [2]. Yet, to train the model to work well on this new antigen, additionally labelled data is needed. Active learning can reduce the number of laboratory tests needed to enhance the model performance in comparison to the random pairing of antigens and antibodies.

1.2 Methods

In active learning, an algorithm can choose which data should be labelled to maximise the performance gain of a model [9][7][1]. The learner (the machine learning algorithm) is trained on a small labelled dataset. It then selects the most useful instance from the unlabelled dataset. This instance is queried to the oracle (e.g. a human annotator) for the label [9]. The aim is to minimise the number of samples that need to be labelled while maximising model performance.

There are many metrics to pick the most informative instance. We investigated the Query-By-Committee and the learning loss method. In Query-By-Committee, the antigen that is causes the most disagreement amongst five models as to the label is selected [10]. Multiple models, known as committee members, are trained on the labelled dataset and used to predict the outcome variable for the unlabelled data instances. The disagreement between the outcomes was measured using the ambiguity metric [3]. In learning loss, the data instance that is expected to have the highest loss is deemed best to label [13]. The target prediction model is supplemented by a loss prediction module that predicts loss for each data instance. Two training methods for this module were tried. It was studied whether these techniques would improve performance of a machine learning model more effectively than the random labelling of antigen.

Antibody-antigen data was generated using the simulation framework Absolut! [8], consisting of 117 antigen mutations and 2230 antibodies. A 8-layer convolutional neural network served as the base machine learning algorithm. This algorithm was initially trained on a small subset of the antibody-antigen data. From the remaining unlabelled data, antigens were selected for labelling either randomly, using Query-By-Committee, or using learning loss.

1.3 Results & Discussion

The active learning techniques did not demonstrate a significant improvement in performance over the random labelling in any of the testing conditions. Learning loss performed slightly better than the random baseline, while Query-By-Committee performed slightly worse. However, these differences were not statistically significant.

For Query-By-Committee, it is suggested that the implemented algorithm fails to pick the most informative antigen to add to the model, but instead trains on outliers, leading to worse performance. Future research could integrate an information density metric in Query-By-Committee to improve performance [4]. The lack of significant results for learning loss could be attributed to the below par accuracy of the loss prediction module. Both loss prediction module training functions led to predicted loss values that did not accurately reflect the true loss values. Alternative training methods should be examined to improve the loss prediction accuracy.

Although active learning techniques did not outperform random labelling in this study, the need for more labelled datasets in antibody-antigen binding will remain. Antigen will continue to mutate, and new antibodies that can be used in antibody drug therapy will continue to be developed. Therefore, continuing research is crucial. Future research should focus on improving the active learning techniques used, but additionally consider explainable alternatives for the convolutional neural network.

References

1. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Machine Learning **15**(2), 201–221 (May 1994). https://doi.org/10.1007/bf00993277, http://dx.doi.org/10.1007/BF00993277
2. Graves, J., Byerly, J., Priego, E., Makkapati, N., Parish, S.V., Medellin, B., Berrondo, M.: A review of deep learning methods for antibodies. Antibodies **9**(2), 12 (4 2020). https://doi.org/10.3390/antib9020012, https://doi.org/10.3390/antib9020012
3. Krogh, A., Vedelsby, J.: Neural network ensembles, cross validation and active learning. In: Proceedings of the 7th International Conference on Neural Information Processing Systems. p. 231–238. NIPS'94, MIT Press, Cambridge, MA, USA (1994)
4. Li, X., Guo, Y.: Adaptive Active Learning for Image Classification. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 859–866 (6 2013). https://doi.org/10.1109/cvpr.2013.116, https://doi.org/10.1109/cvpr.2013.116
5. Lu, R.M., Hwang, Y.C., Liu, I.J., Lee, C.C., Tsai, H.Z., Li, H.J., Wu, H.C.: Development of therapeutic antibodies for the treatment of diseases. Journal of Biomedical Science 2020 27:1 **27**, 1–30 (1 2020). https://doi.org/10.1186/S12929-019-0592-Z, https://jbiomedsci.biomedcentral.com/articles/10.1186/s12929-019-0592-z
6. Pochtovyi, A.A., Kustova, D.D., Siniavin, A.E., Dolzhikova, I.V., Shidlovskaya, E.V., Shpakova, O.G., Vasilchenko, L.A., Glavatskaya, A.A., Kuznetsova, N.A., Iliukhina, A.A., Shelkov, A.Y., Grinkevich, O.M., Komarov, A.G., Logunov, D.Y., Gushchin, V.A., Gintsburg, A.L.: In Vitro Efficacy of Antivirals and Monoclonal Antibodies against SARS-CoV-2 Omicron Lineages XBB.1.9.1, XBB.1.9.3, XBB.1.5, XBB.1.16, XBB.2.4, BQ.1.1.45, CH.1.1, and CL.1. Vaccines **11**(10), 1533 (9 2023). https://doi.org/10.3390/vaccines11101533, https://doi.org/10.3390/vaccines11101533
7. Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Gupta, B.B., Chen, X., Wang, X.: A survey of deep active learning. ACM computing surveys (CSUR) **54**(9), 1–40 (2021)
8. Robert, P.A., Akbar, R., Frank, R., Pavlović, M., Widrich, M., Snapkov, I., Slabodkin, A., Chernigovskaya, M., Scheffer, L., Smorodina, E., Rawat, P., Mehta, B.B., Vu, M.H., Mathisen, I.F., Prósz, A., Abram, K., Olar, A., Miho, E., Haug, D.T.T., Lund-Johansen, F., Hochreiter, S., Haff, I.H., Klambauer, G., Sandve, G.K., Greiff, V.: Unconstrained generation of synthetic antibody–antigen structures to guide machine learning methodology for antibody specificity prediction. Nature Computational Science **2**(12), 845–865 (Dec 2022). https://doi.org/10.1038/s43588-022-00372-4, http://dx.doi.org/10.1038/s43588-022-00372-4
9. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)
10. Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: Proceedings of the fifth annual workshop on Computational learning theory. COLT92, ACM (Jul 1992). https://doi.org/10.1145/130385.130417, http://dx.doi.org/10.1145/130385.130417
11. Taft, J.M., Weber, C.R., Gao, B., Ehling, R.A., Han, J., Frei, L., Metcalfe, S.W., Overath, M.D., Yermanos, A., Kelton, W., Reddy, S.T.: Deep mutational learning predicts ace2 binding and antibody escape to combinatorial mutations in the sars-cov-2 receptor-binding domain. Cell **185**(21), 4008–4022.e14 (Oct 2022). https://doi.org/10.1016/j.cell.2022.08.024, http://dx.doi.org/10.1016/j.cell.2022.08.024

12. Weiner, L.M., Surana, R., Wang, S.: Monoclonal antibodies: versatile platforms for cancer immunotherapy. Nature reviews. Immunology **10**(5), 317–327 (5 2010). https://doi.org/10.1038/nri2744, https://doi.org/10.1038/nri2744
13. Yoo, D., Kweon, I.S.: Learning loss for active learning. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2019)