Paper report

After reading the paper, I learn if independently derived data are very similar to each other, they are called data doppelgängers. When conducting cross-validation, the existence of data doppelgängers may cause models to perform well regardless of how they are trained, which means doppelgänger effects. These data doppelgängers that can confound machine learning are also called functional doppelgängers.

In this paper, the author introduced a variety of data doppelgängers in biological data, including chromatin, protein etc. In fact, there are abundant data doppelgängers in biomedical data. However, I suppose that data doppelgängers also exist in other data. For example, a machine learning model can identify the photos of cats and dogs. In my opinion, there are various species in the family of cat or dog. If some photos in validation set are from the same species as that in training set, although they are from different individuals, the machine learning may perform inflated. Hence, photos of different individuals in the same species can be considered as data doppelgängers.

RCSB Protein Data Bank (PDB)[1] is one of the most widely used databases in my research. After considering carefully, there are three situations in which data doppelgängers occur. (i) Some proteins are able to interact with different small molecule compounds. For example, the NS3 helicase of Zika virus has helicase activity and ATPase activity, it can interact with ssRNA[2], ADP[3] etc. In PDB, we can find different structures of Zika virus NS3 helicase in complex with these molecules respectively (Fig.2 and Fig.3). These structures are very similar to that of Zika virus NS3 helicase[4] (Fig.1). (ii)PDB not only collects wild-type protein structures, some mutant protein structures can also be found(Fig.4 and Fig.5)[5]. These proteins with few mutant residues along with the wild-type protein may be considered as data doppelgängers.(iii) Many homologous proteins are present in biological systems. According to the blood relationship between species, with the proximity of blood relations , their homologous proteins structure are more similar. Therefore, some homologous proteins are also data doppelgängers.

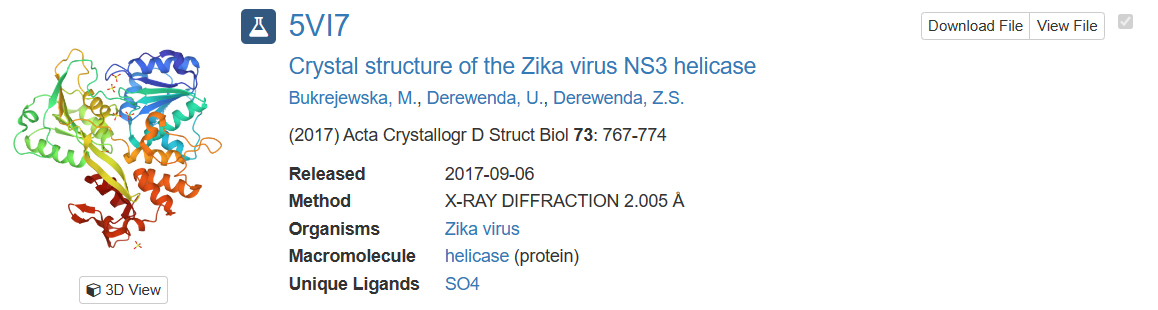


Figure.1

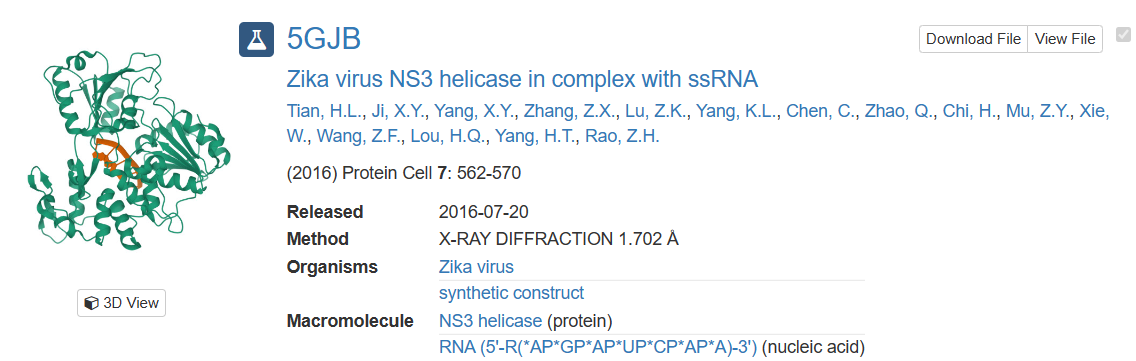


Figure.2

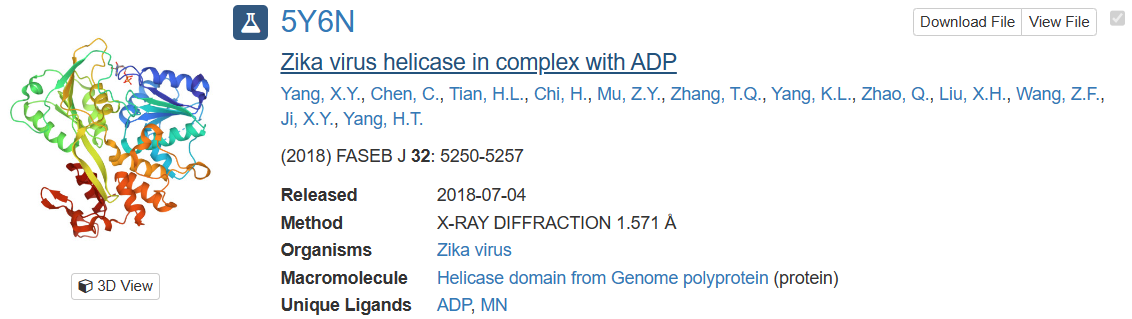


Figure.3



Figure.4



Figure.5

According to the author’s recommendation, it is crucial to be able to identify the presence of data doppelgängers between training and validation sets before validation. How can we identify data doppelgängers in the PDB database? I think a method that compares similarities between proteins can make a difference. In my study of *Structural study of tick-borne encephalitis virus(TBEV) helicase,* I compared the structural of TBEV helicase with the structural of MVEV helicase, ZIKV helicase, DENV-4 helicase, KOKV helicase, JEV helicase, DENV-2 helicase, YFV helicase and KUNV helicase through Dali server[6].The tool calculated RMSD value between TBEV helicase and these proteins(Fig.6), and then build the structural similarity dendrogram(Fig.7). We can find out a cutting off value of RMSD to identify data doppelgängers. Meanwhile, this method needs to be optimized to fit higher throughput.

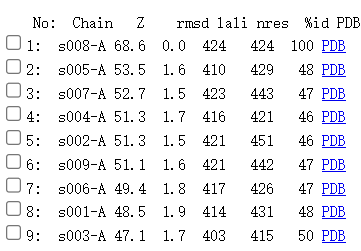


Figure.6 RMSD value

(s001-A represents YFV helicase, s002-A represents DENV-4 helicase, s003-A represents KUNV helicase, s004-A represents KOKV helicase, s005-A represents MVEV helicase, s006-A represents JEV helicase, s007-A represents ZIKV helicase, s008-A represents TBEV helicase, s009-A represents DENV-2 helicase.)

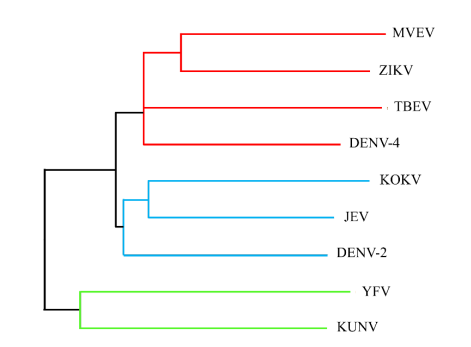


Figure.7 structural similarity dendrogram

[1] https://www.rcsb.org/

[2] Tian HL. et al. Structural basis of Zika virus helicase in recognizing its substrates. *Protein Cell*. 2016 Aug; 7(8): 562–570.

[3] Yang XY. et al. Mechanism of ATP hydrolysis by the Zika virus helicase. *FASEB J*. 2018 Oct;32(10):5250-5257. doi: 10.1096/fj.201701140R. Epub 2018 Apr 17.

[4] Malgorzata B. et al. Crystal structures of the methyltransferase and helicase from the ZIKA 1947 MR766 Uganda strain. *Acta Crystallogr D Struct Biol*. 2017 Sep 1;73(Pt 9):767-774. doi: 10.1107/S2059798317010737. Epub 2017 Aug 15.

[5]Cha HJ. et al. Rescue of deleterious mutations by the compensatory Y30F mutation in ketosteroid isomerase. *Mol Cells*. 2013 Jul;36(1):39-46. doi: 10.1007/s10059-013-0013-1. Epub 2013 Jun 3.

[6] http://ekhidna2.biocenter.helsinki.fi/dali/