数据整理：

原则：数据应该分为3部分，train, val, test.

Train 用来训练模型。

val用来在训练过后查看模型的表现，并对模型进行调整。

Test用来做测试，一般叫做独立测试集，用于评估模型好坏。

Test集应该是完全没有用过的数据集，一般是从别的论文中找到，或者是最新的数据。并且这个test集不应该出现在训练过程中，仅仅用于评估模型，用过就不可以再用了。

所以，test和之前的两个数据集之间是不应该有冗余的。为了有更好的泛化能力，我们也将train对（val + test）做去冗余处理。Train 数据集中的数据是可以有冗余的数据的，因为是点类型的数据。

【许东老师提议】：训练集中随机取数据，针对它对剩下的去冗余，再选第二条。。。由此迭代选择足够的多的验证集。

这个操作做10次，得到10组train集和val集。用作十折交叉验证。（test集另作准备）

【路畅师姐提议】（我当前采用的）

我手头的数据集：全部，3232。0.9：653。 0.3：446。

我们已知446条中是无冗余的。在其中选取45条做test集，选取90条做val集

在3232数据集中，针对这135条选中的数据做CDHIT-0.4（最高只能选到0.4）做去冗余。得到新的train集2558条训练集数据。训练集数据中的蛋白质和val和test分别都没有冗余。

【针对使用被去冗余的数据训练的建议】

思路：被去掉的冗余部分贡献了模型大部分优秀的结果（3232训练效果好，653、446极差）则说明是冗余部分有所贡献。那我们找到冗余那部分蛋白质，关注其特性，找到其特点、分类等信息。之后我们可以对该类蛋白设计网络，对其他类型蛋白设置另一种网络。

同学：没有理由，训练数据中的冗余是需要去掉的。

许东老师：这样的结果可能是因为“上采样”导致的效果好。所以去掉那些蛋白不一定获得更好效果。

实验目的：证明冗余部分蛋白质在模型上不断变好。由此，我们可以挑出这些冗余巨大的蛋白，观察其特性。

流程：数据准备---修改程序---跑实验---统计结果。

设置多组数据：

Data\_all

Data\_1st（去掉446条数据，减少数据冗余少的蛋白的数量）得到2786条蛋白质

【备用数据:1st\_rdc】2786去冗余，假设得到400条数据

Data\_2st（2786去掉400条数据，减少数据冗余少的蛋白的数量）得到2386条蛋白质）

【备用数据:2st\_rdc】2386去冗余，假设得到300条数据

Data\_3st（2386去掉300条数据，减少数据冗余少的蛋白的数量）得到2086条蛋白质）

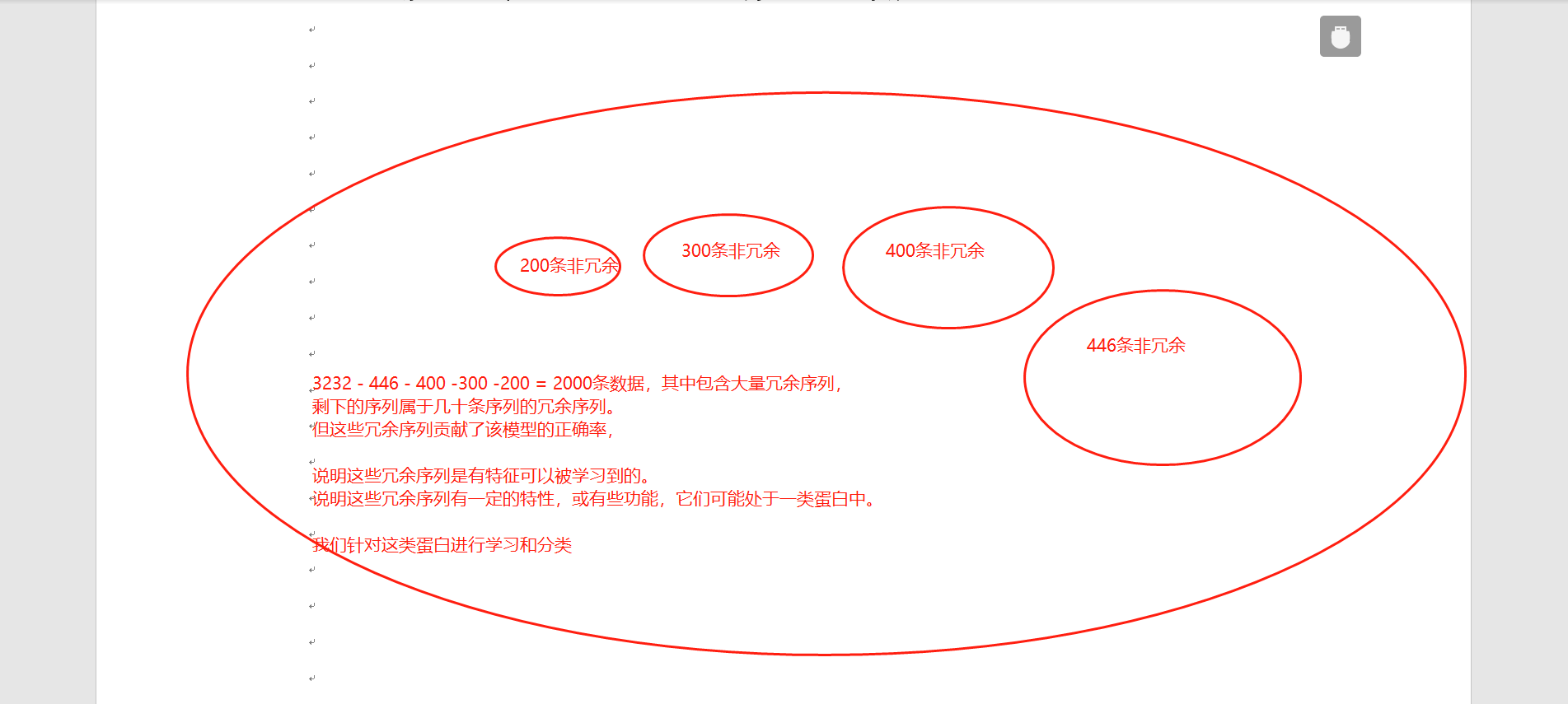
【备用数据:3st\_rdc】2086去冗余，假设得到200条数据

Data\_4st（2086去掉200条数据，减少数据冗余少的蛋白的数量）得到1886约等于2000条蛋白质）

… … …

最后的一组数据中，将包含大量少数几种蛋白的存在冗余的数据。

根据去冗余结果选取3组数据集，运行当前网络，观察效果。



好吧，我先去看看过采样问题。

针对模糊样本的解释与思考：

样本分为正样本、负样本、包含正样本位点的负样本。

第三类样本处于可识别的模糊边界上。模型针对它们的学习很关键。

宫师姐的想法：放后训练。

先用纯正的正样本和负样本进行训练，得到的模型去预测第三类样本。

其中肯定有预测错误的部分。

我们把预测错误的部分单独拿出来，投入模型进行训练。（用以修改那部分预测错误的连线的权重）以此提升模型。

泳冰、静茹反对：这个就像专门拿出错题来做训练，导致的结果会是那些错题都作对了，对正样本预测没啥好处。相当于增加了负样本的预测正确率，并且降低正样本预测准确性。

Attention机制：增加对预测错误样本的训练。有这样的算法和相关研究，需要加紧补充深度学习基础知识。

许东老师建议：这部分数据可以用作模型的预训练，可以考虑不放入正式训练的模型中。并提到王铎林学姐组内正在做相关工作。可以加入帮忙做，但做的足够多才会有论文。

当务之急是找到可以“学得到数据特征的模型”：调整各种超参数，网络层数、网路大小、网络模式、激活函数等等。