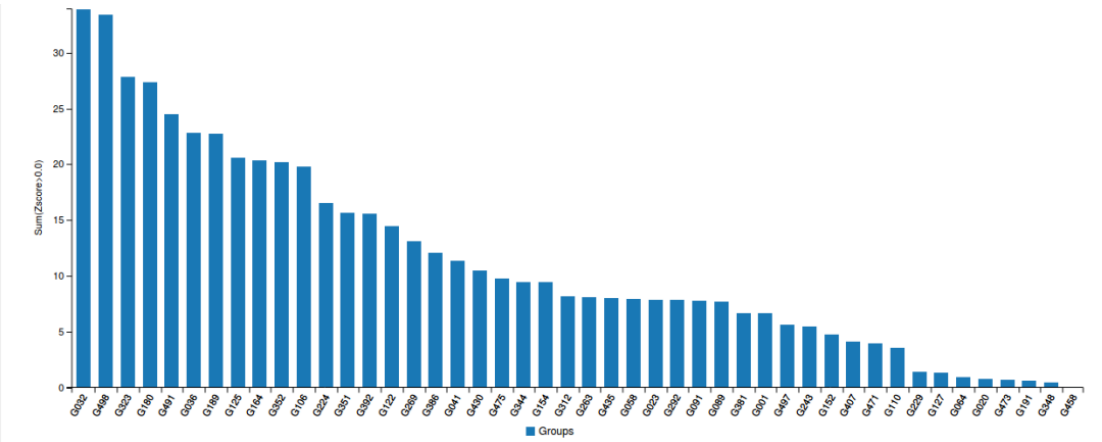
CASP13接触图评测排名第一预测服务器：TripletRes介绍

Original 张贵军（译） iTASSER Today

两年一度的CASP是国际蛋白质结构预测领域的奥林匹克竞赛。深度学习在接触预测领域的深入应用，极大地提升了蛋白质三维结构从头预测的精度。TripletRes是**密西根大学张阳实验室**开发的接触预测服务器，在CASP13接触预测类别46个参赛小组评测中排名第一。



#	GR code	GR name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)	AVG Zscore (>0.0)	Rank AVG Zscore (>0.0)
1	032	TripletRes	31	33.0683	1	1.0667	1	33.9361	1	1.0947	1
2	498	RaptorX-Contact	31	32.9798	2	1.0639	2	33.4022	2	1.0775	2
3	323	TripletRes_AT	30	24.0527	4	0.8684	3	27.8615	3	0.9	
4	180	ResTriplet	31	25.5417	3	0.8239	4	27.3317	4	0.8817	4
5	491	DMP	31	21.1262	5	0.6815	5	24.5195	5	0.7910	5

TripleRes相关论文于2019年12月在《Proteins》发表，英文原文及补充材料参见https://onlinelibrary.wiley.com/doi/10.1002/prot.25798。论文作者李阳和张成辛博士对本中文译稿进行了认真勘校，并提出了宝贵的建议。

Ensembling multiple raw coevolutionaryfeatures with deep residual neural networks for contact-map prediction in CASP13

整合多个共进化原始特征的深度残差神经网络及其在CASP13接触图预测中的应用

作者：Yang Li#，Chengxin Zhang#，Eric W. Bell，Dong-Jun Yu*，Yang Zhang*

第一作者：**李阳，张成辛**
通信作者：**张阳（密西根大学），於东军（南京理工大学）**

摘要：
本文提出了一种纯粹基于共进化特征的新算法，并报道了其在CASP13残基-残基接触项目的预测结果。该算法首先使用两个基于隐马尔可夫模型（HMM）的搜索工具，从多个基因组和宏基因组数据库中收集与目标蛋白序列相似的蛋白序列，构建多序列比对（MSA）；然后，分别基于**协方差**、**逆协方差**和**伪极大似然估计**三种共进化算法分别构建三个特征矩阵，作为深度残差卷积神经网络结构的输入特征，用于接触图训练和预测。通过端到端的训练和叠加，提出了两种集成矩阵特征的整合策略，开发了两个互为补充的接触图预测服务器**TripletRes**和**ResTriplet**。对于CASP13中31个从头结构预测(FM)目标蛋白，TripletRes和ResTriplet两个服务器前L/5远程接触平均预测精度分别为0.640和0.646，其中分别有71%和74%的目标蛋白预测精度高于0.5。进一步分析表明，该算法的优秀性能主要源于灵敏的**MSA构建**方法和先进的**共进化特征整合策略**两个方面，此外结构域分割技术也发现有助于提高接触预测性能。然而，对于尾部区域（通常包含大量的Gap）和同源序列特别少的目标蛋白，接触模型的效果依然不是非常理想。在这些区域和目标蛋白上对模型进行专门训练，可能有助于解决这些问题。

1 绪论

近五十年来，蛋白结构预测算法取得的成功仅限于那些具有同源模板（已测定实验结构）的蛋白质。近年来（译注：尤其是自2014年CASP11以来），得益于基于序列的接触图预测技术的突破，蛋白质结构从头预测方法取得了重大进展。由于接触图在蛋白质结构预测中的重要作用，接触图预测问题受到越来越多研究者的关注，在过去十年中，陆续提出了几种新的预测方法。

蛋白质接触预测的最初研究集中在分析多序列比对（MSA）中两个位置之间的边际相关性。这一思想很有吸引力，但是在接触图预测的实现过程中会引入传递噪声。换句话讲，如果位置A和位置B都耦合到位置C，那么基于边际相关性的分析结果也会显示位置A和位置B之间的存在耦合。随后提出的**直接耦合分析**（DCA）方法（如Onuchic, JN的mfDCA、Jones DT的PSICOV、Söding J的CCMpred和Baker D的

GREMLIN），通过排除其它位置影响来解决传递噪声问题。然而，对于序列同源性较低的蛋白质，由于大多数DCA方法中所使用的Potts逆模型参数不能通过有限的样本数准确估计，导致这类共进化方法不能满足要求。**基于监督机器学习的方法**，如MetaPSICOV（Jones DT，2014）和NeBcon（张阳，2017），通过各种一维序列属性，将多种共进化方法的最终结果合并成特征来预测接触图。这些基于传统监督学习的方法相比于单传使用共进化的算法，在预测精度上有一定提升。最近几年，深度神经网络方法将接触图预测看做像素分类问题来处理，在接触图预测方面取得了巨大的成功（许锦波、程建林等）。然而，仍然有进一步改进的空间，尤其是在特征表达方面。研究发现，大多数机器学习方法都是以共进化分析方法的后处理得分作为特征，因此在后处理过程中可能会出现信息丢失。

与许多其他机器学习预测服务器不同，最近开发的DeepCov（Jones DT，2018）直接使用原始序列协方差矩阵作为其唯一特征，然后使用卷积神经网络来预测接触图，获得了与基于后处理共同进化分析特征的预测器差不多的结果。另外，ResPRE（张阳，2019）考虑协方差逆矩阵的L2正则化估计，能够消除非直接相互作用导致的噪声，当与全残差神经网络结构相结合时，该方法优于其它最先进同类的方法。CASP13期间,我们进一步扩展提出了TripletRes和ResTriplet两种方法，通过两个互补策略，基于端到端训练和堆叠，整合了**协方差**矩阵、**逆协方差**矩阵和伪**极大似然估计的波茨模型参数**矩阵三种原始共进化特征。

在这篇文章中，我们报告了CASP13接触预测项目中TripletRes和ResTriplet的评测结果。我们还仔细分析了预测算法的不同组件的优缺点，并重点研究那些既缺少结构模板又缺少同源序列的FM目标蛋白。还将指出我们方法在CASP竞赛中面临的挑战，以及我们解决这些问题一些可能的思路。

2 材料和方法

TripletRes和ResTriplet整体算法如图1所示。给定查询序列，首先使用DeepMSA（张阳，2019）对多个序列数据库进行增量搜索生成MSA；然后从MSA中提取协方差矩阵（COV）、逆协方差矩阵(PRE)和通过伪极大似然法的波茨模型耦合参数（PLM）三个共进化特征。TripletRes和ResTriplet分别采用了两种不同的策略整合这三种共进化特征。在TripletRes算法中，所有特征都由神经网络直接

融合，实现端到端的训练；在ResTriplet管道中，执行了两阶段的策略，首先从三个特征矩阵中学习产生三个独立的接触图预测结果，然后使用堆叠（stacking）技术将三个独立的接触图预测器整合在一起，辅以二级结构预测特征，输出最终的预测模型。这里，ResTriplet在第一阶段和第二阶段对模型进行单独训练。下面我们将进行详细的介绍。

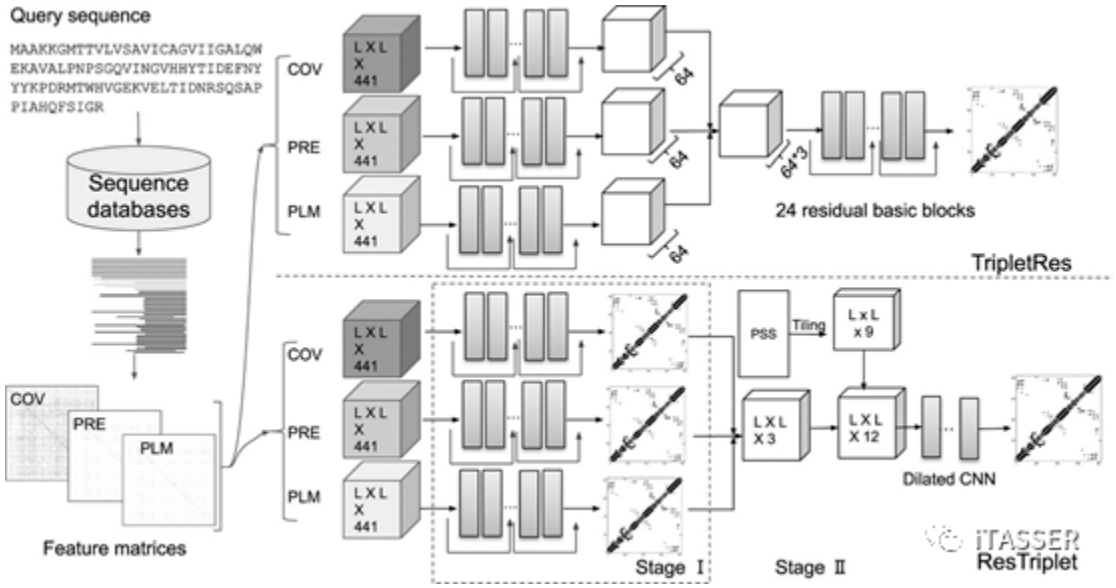


图1：CASP13中TripletRes和ResTriplet接触图预测管道

2.1生成多序列比对（MSA）

MSA是基于共进化分析接触图预测方法的关键步骤。DeepMSA用于从三个序列数据库生成MSA，分三个步骤进行。首先，利用**HHblits对UniClust30**数据库执行最小覆盖率为50%的三次迭代搜索，如果DeepMSA在第1步中没有提供足够的序列(即Nf < 128)，我们将继续进行第2步。Nf为有效序列的规格化个数，其计算方法为:

$$Nf = \frac{1}{\sqrt{L}} \sum_{n=1}^{N} \frac{1}{1 + \sum_{m=1, m \neq n}^{N} \mathbb{I}[S_{n,m} \geqslant 0.8]} \quad (1)$$

其中，L为目标蛋白序列，N为MSA序列数目；如果在MSA中第m条序列和第n条序列的序列相似性大于0.8时，$I[S_{n,m} \geq 0.8] = 1$；否则$II[S_{n,m} \geq 0.8] = 0$。在步骤2

中，使用**Jackhmmer对UniRef90数据库**执行三次迭代搜索，其中E-value设为10。与直接使用Jackhmmer获得的比对序列不同之处在于，我们使用HH-suite的“hhblitdb.pl”脚本从Jackhmmer搜索结果中构建HHblits格式的数据库，并用于进一步的HHblits搜索。如果第2步中MSA的Nf仍然低于128，则执行第3步。在第3步中，使用HMMER软件包中的**HMMbuild工具搜索Metaclust宏基因组序列库**，参数设置为“-E 10 --incE 1e-3”。类似于步骤2，从搜索结果中构建自定义的HHblits数据库。在第2步和第3步中，先前步骤中生成的MSA都被用来对自定义数据库进行HHblits搜索。最终得到的MSA可能非常大，这将导致PLM特征计算的运行时间很长。因此，我们对Nf > 128的MSA删除覆盖率小于60%的同源序列；如果得到的MSA仍然Nf > 128，还可以删除覆盖率小于75%的同源序列。

2.2基于MSA的三种共进化矩阵特征提取

我们的方法使用了三个共进化特征。第一个是由DeepCov（Jones DT等）中提出的协方差矩阵。考虑一个N行L列的MSA，21L x 21L样本协方差矩阵计算如下：

$$S_{ij}^{ab} = f_{i,j}(a,b) - f_i(a)f_j(b) \quad \text{(2)}$$

其中$f_{i,j}(a,b)$是残基类型a和残基类型b在位置为i和j处的观测相对频率；$f_i(a)$是残基类型a在位置i出现的频率。协方差矩阵的每一元素即为残基类型a在i处与残基类型b在j处的协方差。共有21种残基类型（20种标准氨基酸类型加一种Gap类型）。

式(2)中的协方差矩阵蕴含了变量之间的边际相关性，我们通过最小化目标函数来计算第二个特征，即逆协方差矩阵(PRE)。

$$\mathfrak{L} = tr(S\Theta) - log|\Theta| + \rho \|\Theta\|_2^2 \quad \text{(3)}$$

其中在假设数据服从多元高斯分布前提下，前两项可解释为协方差逆矩阵（即精度矩阵Θ）的负对数似然估计。tr(SΘ)为矩阵SΘ的迹，log|Θ| 为矩阵Θ的行列式的对数函数，S为协方差矩阵。式（3）第三项是拟协方差矩阵的L2正则化，ρ设置为e^{-6}。协方差矩阵的逆矩阵剔除了其他位置影响（传递效应），提供了两个残基对之间的直接耦合信息。因此，与协方差矩阵相比，PRE矩阵在接触图的预测方面具

有更好的性能。

负协方差逆矩阵也可解释为逆波茨模型的高斯近似。因此，还考虑了另一种通过伪最大似然估计（PLM）逼近波茨模型的方法。这个过程的出发点是通过观察每个变量的条件概率与所有其他变量的条件概率的乘积来近似序列的概率。这里，我们使用CCMpred计算PLM耦合参数。

协方差矩阵、逆协方差矩阵和波茨模型的耦合参数均采用21L × 21L矩阵的形式，表示在任意两个位置的特定残差类型之间的关系。每一个矩阵的每个位置对都有441个耦合参数的完整集合，可以表示为一个21×21的子矩阵。这样，经过重构，每个序列有三个大小为L×L×441的输入特征。

2.3用于接触模型训练的残差卷积神经网络结构
如图1所示，我们在CASP13中提出了两种基于深度残差神经网络(ResNet)结构，研究接触图预测的最佳集成方式，其中ResNet第一个版本用作基本残差块。这里，每个残差块定义为:

(4)

其中*x*、*y*为所考虑的残差块的输入、输出张量，*f*为激活函数(本文使用ReLU)，*F*为卷积运算学习的残差映射。具体来说，在一个残差块中有两个卷积层。这样，残差函数表示为:

(5)

其中W_1和W_2分别为第一个和第二卷积层的学习权重。为了加快神经网络的训练速度，避免过拟合，我们在基本块中加入了带有默认参数的实例正则化（Instance Normalization）和dropout层。基本残差块的详细结构如图S1所示。这里，dropout率被设置为0.2，这意味着在每个训练批次中，dropout层之前的80%的输入信号将被随机屏蔽。

图S1：基本残差块网络结构

图1的右上半部分描述了TripletRes的网络结构，其中三个共进化的特征由神经网络直接整合。每个输入特征输入到一组24个残差块中，并被转换为64个channels的输出特征。这三个输出特征沿着channel维度串联起来，作为最后一个神经网络的输入。最后一组神经网络通过另外24个残差块从三个变换后的特征中学习模式。所有残差块的channel大小为64，卷积层的核大小设置为3×3，填充大小为1。这样的填充参数设置可以使不同的层保持空间信息不变。在这里，我们使用1×1核大小的卷积层将每个共进化的输入特征和连接的特征转换为64个通道。在这里，我们使用1×1卷积核的卷积层转化每个共进化的输入特征，并将其串接成64个通道。最后

的接触图预测是通过channel设置为1的卷积层输出的sigmoid激活函数获得的。

右下部分展示了一个使用叠加策略的两阶段集成模型ResTriplet。在第一阶段，根据上述三组不同的共进化特征PRE、PLM和COV，分别训练三个单独的基本模型。基本模型具有相同的训练数据和由22个基本残差块组成的相同的神经网络结构。在第二阶段中，我们使用浅层神经网络结构来组合第一阶段的基础模型预测结果，因此，基础模型的预测接触图被视为第二阶段的输入特征。PSIPRED预测的二级结构（如图1中的PSS）作为第二阶段神经网络模型的额外输入特征。对于浅层卷积神经网络，感受野的大小通常是有限的。因此，为了扩大感受野的大小，我们使用了5个扩张的卷积神经网络层，其扩张值设置为2，channel大小设置为16。不同于TripletRes，ResTriplet使用一个扩张卷积神经网络层，其扩张值设置为2，channel设置为1作为最后一层，然后通过在最后一个卷积层上应用sigmoid函数输出。

需要注意的是，在TripletRes和ResTriplet中我们没有对输入特征应用任何类型的预处理操作；而是在每个卷积层之后（除了最有一个卷积层）添加一个实例规范化（Instance normalization）层。TripletRes和ResTriplet采用相同训练集。对于TripletRes，一共训练了10个模型。训练集被分成10个子集，每个子集被视为一个验证集，其余子集被视为每个模型的训练集。最后，所有10个模型的平均值作为输出。在ResTriplet的第一阶段，为了降低过拟合的风险，每个基本模型生成的预测接触图通过10折交叉验证生成。换言之，我们为每个共进化的特征建立10个模型，使用与TripletRes相同的数据分割策略。对于每个特定的共进化特征类型，每个模型验证集的预测接触图作为第二阶段的特征。然而，在第二阶段，由于CASP实验前准备的时间有限，ResTriplet没有进行交叉验证。TripletRes和ResTriplet神经网络基于Pytorch实现，应用Adam优化器训练，初始学习率默认为1e-3，持续50个epochs。TripletRes模型训练需要4个GPU同时运行，而ResTriplet模型训练程序能够只用一个GPU来处理。由于GPU资源有限，在训练过程中考虑了动态批量策略。对用长度L>300的序列，batch大小为1，对用长度L在200~300之间的序列，batch大小为2，对用长度L<200的序列，batch大小为4。出于内存和性能之间的折衷考虑，选择TripletRes和ResTriplet的超参数，特别是层数。虽然更深层次的CNN模型在理论上可以产生更好的性能，但只有有限的层可以放入GPU内存中进

行有效的训练。

2.4结构域分割和基于结构域的接触图预测

结构域是蛋白质结构折叠和功能的基本单位。由于结构域是独立进化的，与全长序列相比，为每个单独的结构域构建MSA并进行单独的预测，通常可以改善结构域内接触预测的准确性。但是，由于事先并不知道CASP目标的结构域边界，对于给定的CASP全长目标，在接触预测之前首先使用ThreaDom（张阳，2013）来识别结构域边界位置。ThreaDom的核心思想是使用LOMETS（张阳，2007）将查询序列穿线到PDB库，构建基于模板结构的多序列比对，随后，计算结构域保守性分数，使用目标特定的得分截止策略确定区域边界。最后从全长序列及其每个结构域推断出接触图预测。结构域之间的接触通过全长序列预测，结构域内接触则通过单个结构域的预测来实现。

3 结果

3.1 整体性能

CASP13发布了90个全长目标，其中82个目标最终结构被公布，评测者将这些目标分成122个结构域。在表S1中，列出了相应目标结构域的Nf值列表，以及ResTriplet和TripletRes在所有 122个结构域中三个范围（短、中、长）内前L、L/2和L/5接触接触预测的准确度，其中L是目标序列的长度。根据CASP官方评测结果，表1汇总了FM目标蛋白中三类短、中、远程接触精度的平均结果。在这里，接触被定义为残基i和j中两个C_β原子间距小于8 Å的。短程接触是指$6 \le |i - j| \le 11$，中程接触是指$12 \le |i - j| \le 23$，远程接触是指$|i - j| \ge 24$。结果表明，两种方法性能相近，TripletRes的预测精度略高于ResTriplet，对于前L和L/5远程接触分别提高了1.2%和0.9%。相应的P值差异为0.82和0.65，表明两种方法在统计学上不显著。考虑前L远程接触精度>0.5的目标数量，TripletRes在31个结构域中有23个域，这也略高于ResTriplet。这种差异可能是由于TripletRes算法是端到端训练的，因此没有受到ResTriplet中堆叠（stacking）策略缺陷的影响。例如，ResTriplet中的特征集合是单独优化的，在一些很难获得同源序列的极端情况下，预测的接触图和二级结构特征可能不太可靠。因此，对于FM目标而言，这种策略的接触精度损失比TripletRes略高。

表1：TripletRes和ResTriple在CASP13 FM目标上的总体性能

注：粗体字体表示在每个类别中性能更好的方法

由于FM目标的平均同源序列数低于基于模板建模（TBM）目标，因此FM目标的接触预测精度预计较低。为了验证这一结论，我们在表S1中列出了所有结构域的接触图预测结果。就我们的情况而言，FM目标和所有目标的平均Nf分别为57.4和390.8。因此，与FM目标相比，ResTriple和TripletRes在所有目标蛋白的前L远程接触平均精度分别高出29%和25%。在所有目标蛋白中，ResTriplet的增加稍微大一点，使得ResTriplet精度都略高于TripletRes。这一现象与FM目标中的趋势正好相反，可能的原因是当存在更多的同源序列时，ResTriplet中二级结构特征信息的预测相对可靠。

3.2 DeepMSA对接触预测精度的影响

TripletRes和ResTriplet直接依赖于从MSA获得的共进化特征，MSA的质量与蛋白质接触图的预测精度密切相关。在CASP13中，我们利用DeepMSA使用多个序列搜索算法跨库搜索来构造MSA。图2A，B显示了使用两种MSA构建机制，一种使用DeepMSA，另一种通过对UniClust30序列数据库进行HHblits搜索的常规方法。TripletRes和ResTriplet对FM目标前L个远程接触预测精度进行逐一比较，进而检验DeepMSA对最终接触预测精度的影响。结果表明，DeepMSA相对于HHblits管道而言，TripletRes和ResTriplet在31个FM目标蛋白中分别有27和28个目标的接触预测精度有所改善。平均而言，DeepMSA将TripletRes/ResTriplet的接触预测精度从33.2%/35.4%提高到40.9%/40.4%。T检验P值为2.9e-06/1.1e-04，这表明改善在统计学上是显著的。由于这两中情况的唯一不同是MSA构建方法，因此差异可单独归因于此因素。在这方面，DeepMSA对31个FM目标蛋白构建MSA的平均Nf为57.4，而HHblits平均Nf为11.6，这表明DeepMSA确实生成了具有更深比对、更多

样的MSA。

图2：MSA对TripletRes和ResTriplet性能的影响。A和B，分别采用DeepMSA的与常规HHblits搜索构建MSA，对TripletRes和ResTriplet的前L远程接触预测结果的影响。C和D，TripletRes和ResTriplet前L/5接触精度与MSA的Nf之间关系。

基于HHblits MSA，在前L远程接触预测精度方面，ResTriplet相比TripletRes略胜一筹；但是两者都使用DeepMSA后，TripletRes比ResTriplet具有更高的接触预测精度。为了更定量地分析MSA对所提出方法性能的影响，我们在图2C和2D中给出了所有目标通过TripletRes和ResTriplet预测的前L/5远程精度与MSA比对深度Nf之间的关系。注意，图中不包括8个实际结构不存在远程接触的目标（即，T0952-D1、T0953s1-D1、T0960-D1、T0960-D4、T0963-D1、T0963-D4、T0979-D1和T0980s2-D1）。在所有122个目标蛋白中，TripletRes和ResTriplet的精度与Nf的常用对数之间的Pearson相关系数分别为0.584和0.551，表明这两个相关性都不是

特别强中度，这主要原因是由于出现了Nf值低但精度高的目标。图2C，2D左上框所示，28个Nf低于10的结构域中，TripletRes和ResTriplet有分别有19和21个结构域接触预测精度超过0.5。为了研究这一现象，采用CCMpred程序对Nf小于10的结构域进行直接耦合分析，结果表明，前L/5远程接触预测精度为0.149，TripletRes和ResTriplet的预测精度为0.591和0.608，与两者相比，CCMpred方法分别降低了74.8%/75.5%。纯DCA方法与基于深度学习的方法在性能上的巨大差距证明了监督深度神经网络训练的有效性。然而，一些MSA比对深度大的目标仍然具有较低的接触预测精度。以T0982-D2为例说明，尽管其Nf值为207.15，但TripletRes和ResTriplet的 L/5远程接触图精度分别为0.462和0.385。根据CASP提供的最好模板(PDB ID: 3tfzB)，该结构域极有可能在体内与特定的配体相互作用。因此，该目标的结构不仅需要**依赖于序列信息**，还需要**依赖于结合配体的信息**，忽略配体信息的会导致接触图预测偏差。另一个可能的原因是在训练过程中使用接触作为类标，本身就是一种粗糙的表示方式，也就是说，二分类的接触图表示可能会导致具体的距离信息的丢失，并且不能真实地表示蛋白质结构的弹性。在前L/5远程接触预测中，由ResTriplet预测的假阳性的平均真值距为10.86，这意味着假阳性仍然足够近，可以相互接触。在这种情况下，二分类接触图训练的模型可能无法区分距离接近接触阈值的残差对，这可能是这种现象的原因。利用距离信息进行训练的研究正在进行中。

虽然DeepMSA提高了平均接触精度，但是对有些目标MSA的过度收集，偶尔也会产生负面的影响。特别是，对于T0982-D2域目标蛋白，DeepMSA通过所有3个步骤来获得最终的多序列比对，并且3个步骤的MSA的Nf值分别为39.7、91.6和288.9。在第1步产生的MSA的基础上，TripletRes和ResTriplet前L/5远程接触精度分别达到了0.962和0.923；当使用步骤2生成的MSA时，ResTriplet的精度略为提高到0.962，而TripletRes的精度则下降到0.615；在第3步的MSA基础上，TripletRes和ResTriplet的精度分别下降到0.577和0.423，并最终分别达到0.462和0.385。随着DeepMSA方法搜索深度的增加，两个预测器的精度呈下降趋势，这表明**更深的MSA不一定会导致更好的接触预测**。部分原因是MSA较深时可能会引入比对噪声，如何定量测定和降低MSA中同源序列的比对噪声仍然是一个值得深入研究的重要问题。

当仅考虑FM目标时，我们发现TripletRes和ResTriplet的Nf值的对数与精度的相关性分别为0.559和0.659。换言之，对于FM目标而言，TripletRes对MSA质量的依赖性较小，这样的观测进一步证明了TripletRes的鲁棒性，尤其是对于FM目标。同时还观察到，在FM目标，ResTriplet的精度与Nf的相关性远高于TripletRes，这证明ResTriplet的性能对MSA的内容比TripletRes更敏感。正如之前指出的，这可能是因为ResTriplet使用了基于MSA的额外的一维预测特征。因此，输入MSA的质量对ResTriplet的最终预测性能有更大的影响。尽管如此，18个FM目标中有11/12的TripletRes/ResTriplet的Nf（<10）非常低，但是却达到了合理的接触精度>0.5。这些数据表明，**即使是从非常有限的同源序列中学习，具有共进化特征的神经网络仍然具有学习潜在接触模式的能力**，这对于缺乏同源序列的FM目标建模来讲是非常重要的。

3.3集成方法及其组件分析

TripletRes和ResTriplet都是综合三个不同组件输入特征信息的混合方法。为了检验信息集成方法的效果，我们将混合方法的结果与三个使用单独输入特征的组件预测器在CASP13　FM目标上进行比较，如图3所示。每个组件预测器的预测值是对应共进化特征类型10个模型的平均值。结果表明，该集成模型对FM目标的影响是非常显著的。例如，TripletRes前L/5的远程接触平均预测精度为64.6%，分别比基于PLM、PRE和COV的组件预测器平均精度高出10.6%、8.6%和12.0%。同样地，堆叠的ResTriplet模型平均预测精度为64.0%，分别比各组件预测器的平均精度高出9.6%、7.6%和10.9%。对前L远程接触预测而言，也能够观察到同样的模式。例如，TripletRes和ResTriplet在CASP“all targets”上前L/5远程接触平均精度分别为75.4%和76.2%，基于PLM、PRE和COV特征的组件预测器分别为75.2%、74.6%和71.0%，TripletRes和ResTriplet在性能上略高于每一个组件预测器。当考虑到前L远程接触预测时也是如此。观测表明，对于FM目标，集成模型是特别必要的。这部分是因为FM目标通常具有较少的同源序列，来自不同特征的互补信息有助于提高最终接触模型的整体精度。然而，对于TBM目标，MSA通常足够深，使得每个组件预测器都能产生令人满意的接触，因此集成模型的效果不太明显。

图3：FM目标上TripletRes和ResTriplet、协方差矩阵特征（COV）、精度矩阵特征（PRE）和逆Potts模型特征（PLM）的耦合矩阵组件预测器的前L和前L/5接触预测的平均精度比较。

此外，在所有的比较中，基于PLM和PRE特征的组件预测器比基于COV特征的组件预测器表现更好。这是因为PLM和PRE特征都是建立在直接耦合分析的基础上的，而COV特征是建立在边际相关分析的基础上的，边际相关分析更容易受到传递噪声的影响。

3.4结构域分割对接触预测的影响

TripletRes和ResTriplet都使用DeepMSA从单独结构域序列中创建MSA，每个结构域则是通过ThreaDom从全长蛋白解析分割得到。为了检验结构域分割对接触预测的影响，图4对比了ThreaDom在26个全长蛋白作为多结构域序列进行结构域分割前后的前L远程接触预测精度。在这26个蛋白中，有59个结构域被公布，图4比较了被ThreaDom指定为多结构域序列的26个全长蛋白，在结构域分割前后的前L远程接触预测精度。在这26个蛋白中，有59个结构域被公布，图4列出了这59个结构域的接触预测结果，这些分割结果与CASP13的官方定义结构域保持一致。

图4：结构域划分和无结构域划分的前L远程接触预测精度比较。A、TripletRes；B，ResTriplet

基于ThreaDom的结构域划分对接触预测产生了明显的积极影响。在59个结构域中，经过结构域划分过程，TripletRes(或ResTriplet)有23个(或36个)结构域的预测精度高于基于全链的接触预测精度，有12个(或7个)结构域的预测精度低于基于全链的接触预测精度。平均而言，TripletRes和ResTriplet的前L远程接触预测精度（52.6%和54.1%）也高于全链预测精度(46.9%和46.6%)，相应的P-value为1.8e-03和9.9e-05，表明差异具有统计学意义。

在图4中的59个结构域中，有两个结构域，分别为T0981-D3和T0981-D5，在使用和不使用域划分之后预测精度存在着非常大的差异。T0981它包含五个结构域，ThreaDom将目标序列划分成四个域(参见图S3)，T0981-D3和T0981-D5两个结构域都来自T0981。进一步观察发现，当使用全链序列预测时，MSA中的标准化有效序列数（Nf）为0.2，MSA中仅存在9条同源序列。经过ThreaDom的结构域划分之后，T0981-D3和T0981-D5的Nf值分别增加到229.6和2.8，最终将T0981-D3的TripletRes/ResTriplet接触预测精度从0.187/0.177提高到0.675/0.818，将T0981-D5的接触预测精度从0.110/0.244提高到0.803/0.669。这些数据表明，结构域划分的优势主要体现在MSA的改进上，结构域划分有助于DeepMSA为每个单独的结构域检测到更多的同源序列。

然而，从理论上讲，结构域划分也可能导致DCA模型估计的偏差。在结构域划分之前，DCA模型中某一氨基酸在某一位置的概率取决于全长序列的所有其他位置。然

而，在结构域划分之后，这个概率只取决于其所在结构域内的其它位置。此外，结构域划分后，靠近结构域边界位置的比对质量也会受到负面影响。这两个因素可能就是少数结构域中MSA即使存在许多的同源序列，但预测精度也较低的原因。

3.5哪些策略是有帮助的

我们发现，使用原始的**共进化特征**，结合**深度卷积神经网络**，可以提高接触图预测精度。在共进化特征中，基于DCA的特征（即逆协方差矩阵和伪最大似然估计的波茨模型参数）表现优于边际相关特征（即协方差矩阵）。考虑到现有文献中许多方法都使用了协方差矩阵特征，这一结论可能有助于进一步扩展接触图预测的研究方向。此外，尽管单个原始共进化特征可以提供相对准确的接触图预测精度，但是基于深度卷积神经网络的多特征融合/集成仍然会获得更好的性能。

根据我们的自己的测试基准和CASP13的结果，多种序列比对生成协议(包括搜索算法和序列数据库)的组合可以显著提高接触预测精度，特别是对FM目标而言。我们预测过程的另一个重要方面是结构域划分技术，即使在预测的结构域边界不精确的情况下，结构域划分仍然可以通过提供更多样和更深的MSA来提高精度。

3.6出现了什么问题

由于所有的进化耦合都是从MSA中推断而来的，在MSA中跨度长的Gap会导致假阳性耦合信号，这意味着末端区域可能会引入强噪声，因为Gap在构建共进化特征时被视为一种（额外）的氨基酸类型。这个问题在T0957s2-D1目标中表现的尤为严重，TripletRes和ResTriplet的前L远程接触精度分别为39.4%和34.2%。如图5A所示，在前30个残基中存在着大量的比对Gap，因此，大多数假阳性预测（在图5B中的接触图左上角标记为灰色圆圈）来自涉及这些N-末端残基的接触。图5C展示了该区域的三维结构，它显示了N-端与结构域的其余部分很好地堆积在一起，并且具有与其他残基相同的二级结构，这表明这些间隙不是由于不规则的局部结构或无序模体（motif）造成的。在这方面，如何适当考虑MSA中的较大Gap仍是一个重要问题。

图5：CASP13中T0957s-D1结构域示例。由于比对中的Gap数目较多，在N端尾部区域中出现假阳性接触预测。A、沿查询序列的Gap数条形图。B、实际结构（右下三角形部分）的接触与ResTriplet（左上部分）预测的接触点，其中灰色圆圈和黑色正方形分别表示假和真阳性预测。C、T0957s-D1的三维实验结构，N端尾部用黑色标记。

4结论

我们介绍了TripletRes和ResTriplet两种混合的接触预测方法，并在CASP13中进行了测试。与其它使用后处理的共进化分析耦合势能方法不同，这两种方法使用原始的共进化矩阵作为唯一的输入特征，当与深度残差神经网络相结合时，可以形成先进的接触图预测方法。这些方法的成功部分归功于特征集成策略：用神经网络进行特征融合，称为TripletRes；用多个预测器堆叠，称为ResTriplet。同时，DeepMSA迭代的MSA构建过程，结合多个序列数据库的来源，以及结构域特定MSA集合都有助于提高最终的接触图预测性能。这些方法的效果和实用性在FM目标上特别明显，与较简单的TBM目标相比，FM目标通常涉及到较少的同源序列。

然而，在接触图预测方面仍然存在一些问题，特别是在序列的尾部区域通常有较多的比对Gap，导致接触精度低于其他区域。同时，对于同源序列较少的硬目标，接触预测精度还远远不能令人满意。开发新的管道，采用专门针对这些尾部区域和难目标的模型，将可能有助于解决这些问题。

致谢：
我们感谢郑伟博士有指导性的讨论。使用XSEDE（由美国国家科学基金会（ACI-1548562）资助）完成对TripletRes和ResTriplet模型的训练。这项工作是李阳博士生访问密歇根大学张阳实验室时完成的。

原文链接：https://onlinelibrary.wiley.com/doi/10.1002/prot.25798