Computer Science and Application 计算机科学与应用, 2022, 12(11), 2417-2423
Published Online November 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.1211247

Hans汉斯

基于深度学习的预测TCR和肽的相互作用算法研究

王　宁，马　欣*

天津工业大学，天津

收稿日期：2022年9月30日；录用日期：2022年10月28日；发布日期：2022年11月4日

摘　要

免疫疗法是一种利用人体自身免疫系统反应的癌症治疗方法。继手术、化疗和放疗等传统治疗方法后，免疫疗法逐渐成为了最有前途的癌症治疗方法。在各种免疫疗法中，发展较快的是TCR-T疗法。TCR-T疗法通过基因编辑技术，把特异性识别肿瘤抗原肽的T细胞受体(TCR)基因导入到患者自身的T细胞，使患者的T细胞能够表达外源性TCR，并且获得特异性杀伤肿瘤细胞的能力。然而，如何从大量非结合TCR中筛选出特异性识别肿瘤抗原肽的TCR，这一直是免疫和生物信息学领域的挑战。为解决这一问题，本文提出了一种基于深度学习的分类算法，对TCR与多肽的相互作用进行准确预测。实验结果表明，本文提出算法在数据集上AUC的值为0.8513。此外，还与随机森林和NetTCR分类模型进行不同的分类指标对比，该算法的指标值都有着较高的提升。

关键词

免疫疗法，深度学习，TCR，肽，相互作用

Research on Predicting TCR and Peptide Interaction Algorithms Based on Deep Learning

Ning Wang, Xin Ma*

Tiangong University, Tianjin

Received: Sep. 30th, 2022; accepted: Oct. 28th, 2022; published: Nov. 4th, 2022

Abstract

Immunotherapy is a type of cancer treatment that exploits the body's own immune system response.

*通讯作者。

文章引用: 王宁, 马欣. 基于深度学习的预测 TCR 和肽的相互作用算法研究[J]. 计算机科学与应用, 2022, 12(11): 2417-2423. DOI: 10.12677/csa.2022.1211247

Immunotherapy is emerging as the most promising cancer treatment after traditional treatments such as surgery, chemotherapy and radiotherapy. Among the various immunotherapies, TCR-T therapy develops rapidly. TCR-T therapy introduces the T Cell Receptor (TCR) gene that specifically recognizes tumor antigen peptide into the patient's T cells by gene editing technology, so that the patient's T cells can express exogenous TCR and obtain the ability to specifically kill tumor cells. However, how to screen out TCRs that specifically recognize tumor antigenic peptides from a large number of unbound TCRs has been a challenge in the fields of immunity and bioinformatics. To solve this problem, this paper proposes a classification algorithm based on deep learning to accurately predict the interaction between TCR and peptides. The experimental results show that this paper proposes that the algorithm has a value of AUC of 0.8513 on the dataset. In addition, there are also different classification indexes compared with the random forest and NetTCR classification models, and the index value of this algorithm is highly improved.

Keywords

Immunotherapy, Deep Learning, TCR, Peptide, Interaction

1. 引言

近几年来，随着基因组学和蛋白质组学的不断发展以及医学前沿技术的迅速突破，免疫疗法作为多种癌症的治疗方法不但获得了巨大的成功，并且具有良好的应用前景。免疫疗法是一种利用人体自身免疫系统来对抗癌症的方法，对癌症治疗领域的发展至关重要。随着全球免疫学、医学和细胞生物学等相关领域的科学研究工作不断进步，全球免疫疗法的研究进展也正在快速提升，并且已经取得很多研究成果。继手术、化疗和放疗等传统治疗方法后，免疫疗法技术已经逐渐成为了治疗癌症领域的新兴方法。目前，发展较为迅速的免疫疗法主要包括 TCR-T (T Cell Receptor-Gene Engineered T Cells)和 CAR-T (Chimeric Antigen Receptor T Cells)等疗法[1]。CAR-T 疗法对血液肿瘤疾病取得良好的治疗效果，然而 CAR-T 免疫疗法存在着一些缺点。首先是它对肿瘤细胞的识别机制，CAR-T 是通过基因编辑技术人工设计的单链抗体片段(CAR)，该片段对识别癌细胞膜表面表达的抗原较为有效，然而癌细胞内部还存在着大量的实体瘤抗原，CAR-T 细胞并不能对胞内抗原进行识别。因此，CAR-T 疗法对实体肿瘤的治疗效果并不十分理想。然而，通过 TCR-T 疗法设计的带有编辑过 TCR 的 T 细胞不但可以识别癌细胞表面的抗原，而且还能识别 MHC 递呈的癌细胞内部抗原。TCR-T 细胞上的 TCR 通过结合 MHC 分子递呈出来的抗原肽段识别癌细胞，进而主动攻击将癌细胞杀死，达到治疗癌症的目的。

2. 现有研究技术

2.1. TCR 和肽的相互作用原理

TCR-T 细胞是通过自身表面的 TCR 识别 MHC 分子递呈的抗原肽而产生相互作用的，TCR-T 细胞上的 TCR 与 MHC I 类分子呈递的肽结合的过程如图 1 所示[2]。抗原呈递细胞(APC)内的抗原蛋白质分子首先被蛋白质裂解酶水解为大量的多肽。接着，多肽通过抗原加工相关的转运蛋白的转运，输送到内质网中，MHC I 类分子与内质网上的多肽结合后，呈递到细胞膜表面。最后，TCR 识别 MHC Ⅰ分子呈递的多肽，两者结合之后产生胞间信号，T 细胞和肿瘤细胞之间出现免疫应答反应，进而杀死癌细

胞，达到治疗肿瘤的目的。然而，并不是所有的 MHC 分子所呈现的多肽都能和 TCR 发生相互作用，其中，很多多肽片段不具有免疫原性，也就不会被 TCR 所识别。因此，如何从大量非结合 TCR 中筛选出特异性识别肿瘤抗原肽的 TCR，成为了目前 TCR-T 疗法中一大挑战。通过高通量设备和实验方法，从大量的 TCR 和候选肽中筛选出可以相互作用的配对序列显然是不现实的。这样做不仅耗费大量的时间，而且需要很多资金支持。如何快速准确地预测 TCR 和肽之间的相互作用已经成为近些年来生物医学领域重点研究方向。

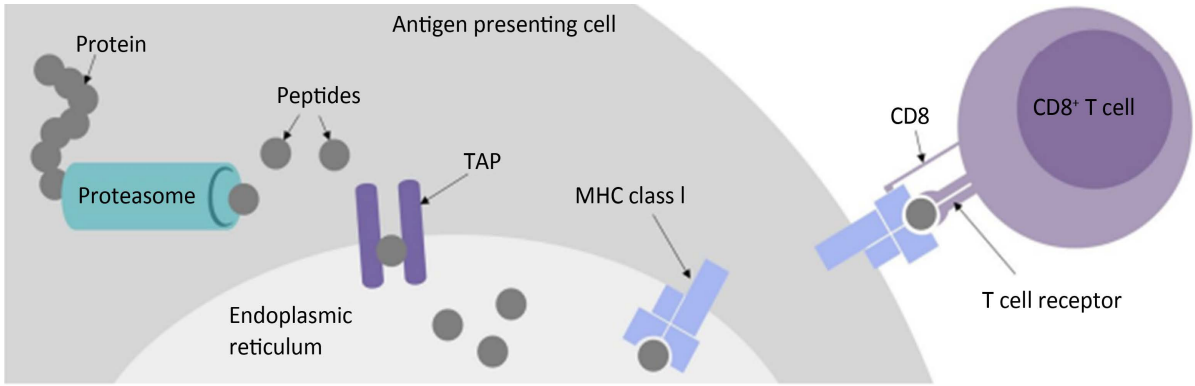

Figure 1. The process of the TCR and peptide interaction
图 1. TCR 和肽的相互作用的过程

2.2. 现有研究方法

随着人工智能技术的不断发展，生物医学与人工智能的结合逐渐成为新的突破点。人工智能领域的机器学习和深度学习技术作为一种强有力的方法为预测 TCR 和肽之间相互作用提供了指导。到目前为止，现有的预测 TCR-肽的相互作用的方法主要基于分子动力学、机器学习和深度学习三类。在预测 TCR 和肽结合的亲和力的早期研究当中，Michielin 等人利用热力学积分法计算自由能，基于分子动力学分析结合的 TCR 和肽之间的自由能的差异，来反映 TCR 和肽之间结合亲和力的强弱[3]。该方法对于预测 TCR 和肽亲和力具有重大的生物意义。2019 年，Gielis 等人研发了一种基于机器学习的随机森林分类模型 TCRex 预测 TCR 和肽的相互作用[4]。该模型从 McPAS-TCR 数据库和 VDJdb 数据库中筛选出 TCR 的 β 链序列和配对的肽序列，训练和评估 TCRex 模型。该分类器使用 5 倍分层交叉验证方法计算了每个表位特异性模型的 ROC 曲线下面积(AUC)值。模型的 AUC 至少为 0.7，体现出该模型较强的泛化能力。Jurtz 等人在深度学习的基础上设计了一种基于卷积神经网络的分类模型 NetTCR [5]，该模型能够预测 TCR 和 MHC I 等位基因 HLA-A*02:01 呈现的肽之间的相互作用。NetTCR 模型依赖于肽的氨基酸序列和 TCR β 链的 CDR3 区域作为输入，这两个序列都是使用 BLOSUM50 矩阵进行特征表示。CNN 非常适合处理长度不同的未对齐肽和 TCR 序列，使用卷积滤波器扫描两个输入，CNN 能够整合整个输入序列中不同过滤器发现的模式的信息。接着对模型进行训练和评估。结果表明，该模型可以从大量非结合 TCR 中识别结合给定同源 MHC 呈递的肽靶点的 TCR。

然而这些现有的研究方法虽然能够对 TCR 和肽的相互作用进行预测，但是由于 TCR-肽序列配对的多样性和可用训练数据有限，这些模型对于已经训练和测试过的数据预测效果较好，然而对未覆盖的或潜在的肽或 TCR 序列的新数据预测能力一般，模型的泛化能力和稳定性较差[6]。因此，需要寻找一个较好的特征工程方案对 TCR 和肽序列数据进行特征编码，并且开发出一种基于深度学习的高性能和高稳定性的方法，这对预测 TCR 和肽的相互作用至关重要[7]。

3. 基于深度学习的 TCR 和肽的相互作用预测模型

基于上述预测 TCR 和肽的相互作用模型的研究现状和存在的问题，本文从两个不同角度分别对 TCR 和肽进行数值化表征，提出了一种基于深度学习的神经网络模型(C-L-M)预测 TCR 和肽的相互作用。C-L-M 模型从不同的数据挖掘角度出发，不仅为TCR 和肽的相互作用预测提供了新的认知，而且为TCR-T 免疫治疗方法中 TCR 和肽的初步筛选提供了一定的指导作用。

3.1. 实验数据集

本实验所有的数据都是从 TCRdb 数据库中经过一定条件筛选出来的，一共包含 3416 对 TCR 和肽相互配对的阳性数据和 3416 对的阴性数据，一共包含 6832 对数据，这些数据都是经过实验验证的。数据集按 7:2:1 划分为训练集、验证集和测试集。本研究所用数据集的分布如表 1 所示。

Table 1. Distribution of the datasets used in this paper
表 1. 本文所用数据集的分布

性质	训练集	验证集	测试集
阳性数据	2390 对	684 对	342 对
阴性数据	2390 对	684 对	342 对
数据总量	4780 对	1368 对	684 对

3.2. 数据的预处理和特征表示

对 TCR 和肽的氨基酸序列构建一个有效的特征数值化方法，将会对预测两者相互作用结果的准确率和精准度造成直接影响。本研究首先对人类的二十种氨基酸用自定义的 getDict 函数定义了一个字典，把 TCR 和肽的氨基酸序列作为一个整体(TCR-肽序列)，使用自定义的 toToken 函数把字符序列转换为数值序列。接着从不同角度对 TCR-肽数值序列进行特征提取。一个角度对 TCR-肽数值序列使用 One-Hot 编码方式表示[8]，另一个角度是对 TCR-肽数值序列用 Embedding 编码方式来表示[9]。通过对 TCR-肽序列从两种从不同角度分别进行特征提取，把这两种表示方法相结合，对 TCR 和肽的相互作用预测进行特征表示。

3.3. C-L-M 模型的构建

由于一维卷积神经网络(Conv1D)和循环神经网络(LSTM)在处理文本序列具有很好的性能，而多层感知机(MLP)具有处理高维、复杂的数据的能力。因此，我们把 Conv1D、LSTM 和 MLP 进行了一个组合，设计了一种基于深度学习的模型 C-L-M。让 Conv1D 去处理 One-Hot 编码方式表示的 TCR-肽序列的特征矩阵，让 LSTM 去处理用 Embedding 编码方式来表示的 TCR-肽序列的特征矩阵。最后，用 Concatenate() 函数把两个输出向量的特征进行融合，输入到设计的 MLP 神经网络中，对 TCR 和肽的相互作用进行预测。基于深度学习的预测 TCR 和肽的相互作用的 C-L-M 算法流程图如图 2 所示。

3.4. 模型的训练和超参数设置

本研究把 One-Hot 编码和 Embedding 编码 TCR-肽序列分别经过 Conv1D 和 LSTM 进行处理。Conv1D 神经网络是由 2 个卷积层、2 个池化层和 1 个全连接层构成，卷积核数量都是 64，全连接神经元数量为 128。LSTM 神经网络是由 2 个 LSTM 和一个全连接层构成，神经元个数分别为，64、64 和 128。随后，把两个输出向量用 Concatenate()函数进行特征融合，输入到具有 3 个全连接层的 MLP 神经网络中，神经

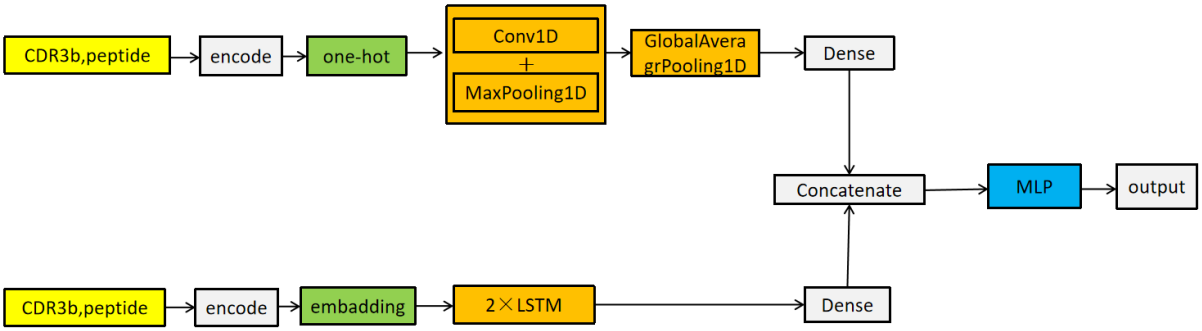

Figure 2. Flow chart of the C-L-M algorithm based on deep learning
图 2. 基于深度学习的 C-L-M 算法流程图

元个数分别为 128、64，和 32。Conv1D、LSTM 和 MLP 中的隐藏层全部采用采用 ReLU 函数作为激活函数。在全连接层之后加入 Dropout 层，rate 设置为 0.2。加入 Dropout 层的作用是防止网络过拟合，随机让一定数量的神经元停止工作，这样可以提高网络的泛化能力[10]。最后，加一个全连接层作为输出层，神经元个数为 1，Sigmoid 作为激活函数(表达式如式(2)所示)，输出结果。本研究使用二值交叉熵(Binary_crossentropy)作为 C-L-M 模型的损失函数(表达式如式(3)所示)，损失函数的目的是为了计算 TCR 和肽的相互作用的概率值和实际值之间的误差，其中，y_i 为真实值，$\hat{y}_i$ 为预测值。本研究使用的优化器是 RMSprop 算法(Root Mean Square prop)，学习率为 0.001。使用 RMSProp 算法对神经网络中神经元的权重 W 和偏置 b 的梯度进行更新[11]。其中，ReLU、Sigmoid 和 Binary_crossentropy 的公式如下：

$$F(x)=\max(0,x) \tag{1}$$

$$S(x)=\frac{1}{1+e^{-x}} \tag{2}$$

$$\text{LOSS}=-\frac{1}{\text{output size}}\sum_{i=1}^{\text{output size}} y_i\cdot\log\hat{y}_i+(1-y_i)\cdot\log(1-\hat{y}_i) \tag{3}$$

4. 实验结果与分析

4.1. 实验环境

在软件环境方面，操作系统为 Windows 10，Python 环境版本为 Python 3.9，使用的框架版本是 Keras 2.5 和 Tensorflow 2.5。在硬件环境方面，使用的 CPU 是 AMD Ryzen 7 4800U with Radeon Graphics 1.80 GHz，计算机的内存是 16G，使用的显卡是 GTX 1060。

4.2. 性能指标

把预测 TCR 和肽的相互作用的模型构建完成后，我们需要对模型的性能进行一个评估。不同种类的模型具有不同的评估指标，这些指标能够客观的评估模型性能的好坏。选择合适的指标可以帮助我们对模型的性能进行调整。通过把不同的指标进行比较，从各种模型中选择最优的一种作为最终模型。本研究是对 TCR 和肽的相互作用的预测结果进行一个分类，因此，采用分类性能的评价指标作为评估模型的标准。本研究选择了 ROC 曲线、Precision、Recall 和 AUC 作为 C-L-M 模型评估的指标。其中，TP 代表真阳性，即正确识别 TCR 与配对肽结合；TN 代表真阴性，即正确识别 TCR 不结合配对肽；FN 代表假阴性，错误识别 TCR 不结合配对肽。Precision 和 Recall 的公式如下：

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

4.3. 实验结果和分析

为了评估 C-L-M 模型的性能，我们将 C-L-M 模型与基于机器学习的随机森林分类模型和 Jurtz 等人提出的基于卷积神经网络的 NetTCR 分类模型在本文研究的数据上进行对比。C-L-M 模型与其他两个模型的 ROC 曲线对比如图 3 所示。由此图可以看出，C-L-M 模型的 ROC 曲线相对与随机森林和 NetTCR 更加靠近左上角，这表明 C-L-M 模型要比其他两个的预测效果都要好。而 C-L-M 模型与其他模型在其他指标的对比结果如表 2 所示，从该表我们得知，C-L-M 模型的 AUC 为 0.8513，而随机森林分类模型和 NetTCR 分类模型在 AUC 方面表现一般，分别为 0.8263 和 0.8039。与随机森林和 NetTCR 模型相比，C-L-M 模型在 AUC 方面分别提高了 0.025 和 0.0474。在其他指标方面，与 NetTCR 相比，C-L-M 模型的 Precision 提升了 0.0489。与随机森林相比，C-L-M 模型的 Recall 提高了 0.0761。因此，C-L-M 模型在预测 TCR 和肽的相互作用的性能有着不错的表现。

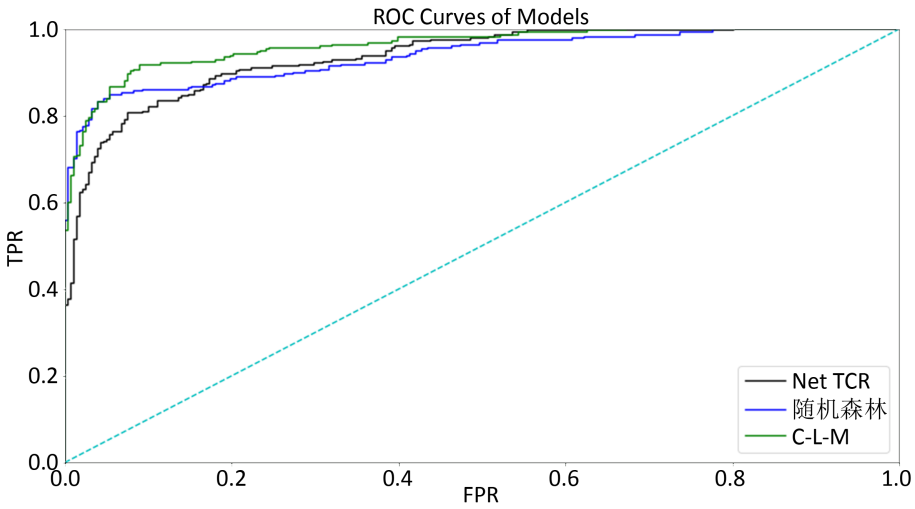

Figure 3. ROC curve comparison of C-L-M and other classification models
图 3. C-L-M 与其他分类模型的 ROC 曲线对比图

Table 2. Comparison of various indexes of C-L-M and other classification models
表 2. C-L-M 与其他分类模型的各种指标对比

Model	Precision	Recall	AUC
随机森林	0.7595	0.9102	0.8263
NetTCR	0.7302	0.9641	0.8039
C-L-M	0.7791	0.9863	0.8513

5. 结论

深度学习形式的人工智能和免疫生物学的结合对于进一步了解 TCR 和肽的相互作用至关重要。为了从大量 TCR 和肽的氨基酸序列中快速地、高通量地筛选出具有高亲和力和稳定性的 TCR-肽对序列，本文提出了一种基于深度学习的分类模型 C-L-M，该模型从两个不同的角度对 TCR-肽序列进行特征提取，并把 Conv1D、LSTM 和 MLP 神经网络进行组合，实现 TCR 和肽的相互作用预测。实验结果显示，与其

他分类模型相比，C-L-M 模型在特征提取方面和处理复杂数据方面性能表现更好。该方法不仅提高了 TCR-T 疗法中 TCR 和肽的筛选效率，而且对细胞免疫治疗方法的研究起到了推动的作用。

参考文献

[1] Zhang, C., Liu, J., Zhong, J.F. and Zhang, X. (2017) Engineering CAR-T Cells. *Biomarker Research*, **5**, Article No. 22.

[2] Mösch, A., Raffegerst, S., Weis, M., Schendel, D.J. and Frishman, D. (2019) Machine Learning for Cancer Immunotherapies Based on Epitope Recognition by T Cell Receptors. *Frontiers in Genetics*, **10**, Article 1141. https://doi.org/10.3389/fgene.2019.01141

[3] Michielin, O. and Karplus, M. (2002) Binding Free Energy Differences in a TCR—Peptide-MHC Complex Induced by a Peptide Mutation: A Simulation Analysis. *Journal of Molecular Biology*, **324**, 547-569. https://doi.org/10.1016/S0022-2836(02)00880-X

[4] Gielis, S., Moris, P., Bittremieux, W., *et al*. (2019) TCRex: Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires. *BioRxiv*, Article ID: 373472. https://doi.org/10.1101/373472

[5] Jurtz, V.I., Jessen, L.E., Bentzen, A.K., *et al*. (2018) NetTCR: Sequence-Based Prediction of TCR Binding to Peptide-MHC Complexes Using Convolutional Neural Networks. *BioRxiv*, Article ID: 433706. https://doi.org/10.1101/433706

[6] Springer, I., Besser, H., Tickotsky-Moskovitz, N., Dvorkin, S. and Louzoun, Y. (2020) Prediction of Specific TCR-Peptide Binding from Large Dictionaries of TCR-Peptide Pairs. *Frontiers in Immunology*, **11**, Article 1803. https://doi.org/10.3389/fimmu.2020.01803

[7] Akbar, R., Robert, P.A., Pavlović, M., *et al*. (2021) A Compact Vocabulary of Paratope-Epitope Interactions Enables Predictability of Antibody-Antigen Binding. *Cell Reports*, **34**, Article ID: 108856. https://doi.org/10.1016/j.celrep.2021.108856

[8] Rodríguez, P., Bautista, M.A., Gonzàleza, J. and Escaleraac, S. (2018) Beyond One-Hot Encoding: Lower Dimensional Target Embedding. *Image and Vision Computing*, **75**, 21-31. https://doi.org/10.1016/j.imavis.2018.04.004

[9] Cellucci, C.J., Albano, A.M. and Rapp, P.E. (2003) Comparative Study of Embedding Methods. *Physical Review E*, **67**, Article ID: 066210. https://doi.org/10.1103/PhysRevE.67.066210

[10] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research*, **15**, 1929-1958.

[11] Kurbiel, T. and Khaleghian, S. (2017) Training of Deep Neural Networks Based on Distance Measures Using RMSProp. *ArXiv*, 1708.01911.