南京农业大学大学生创新训练项目计划申请书

项目编号_____

项目名称 基于SeqGAN模型生成芬太尼类衍生物

项目负责人 陈述 **联系电话** 13905663645

所在学院 理学院

学号 9203012003 **专业班级** 统计 201 班

指导教师 骈聪

E-mail piancong@njau.edu.cn

申请日期 2022年3月22日

项目期限 2022 年 4 月-2023 年 4 月

南京农业大学 教务处

南京農業大學
NANJING AGRICULTURAL UNIVERSITY

填写说明

1. 本申请书所列各项内容均须实事求是，认真填写，表达明确严谨，简明扼要。

2. 申请人可以是个人，也可为创新团队，首页只填负责人。“项目编号”一栏不填。项目团队一般不超过 5 个人。

3. 本申请书为大 16 开本（A4），左侧装订成册。可网上下载、自行复印或加页，但格式、 内容、大小均须与原件一致。

4. 申请国家级、省级项目的负责人提交的申请书经所在学院认真审核，经初评和答辩，签署意见后，将申请书（ 一式两份）报送南京农业大学教务处实践科。

5. “经费预算”主要包括：元器件费、实验耗材费、测试化验加工费、资料费、学术交流会议费、调研费、在学术期刊发表论文的版面费、知识产权事务费、市内交通费、其他与项目研究有关的费用等。不得包括下列事项相关费用：餐饮、旅游、通讯、计算机及其配件、设备维修、办公耗材（墨盒、硒鼓、打印纸等）、劳务费、其它与项目研究无关的事务。

南京農業大學
NANJING AGRICULTURAL UNIVERSITY

一、 基本情况

<table>
<tr><td>项目名称</td><td colspan="7">基于SeqGAN模型生成芬太尼类衍生物</td></tr>
<tr><td>所属学科</td><td colspan="7">学科一级门： 数学　　学科二级类： 统计学</td></tr>
<tr><td>项目来源</td><td colspan="7">○ 学生自主选题， 来源于自己对课题的长期积累与兴
◉ 趣 B、 学生来源于教师科研项目选题
○ C、 学生承担社会、 企业委托项目选题
○ D、拔尖专项
○ E、竞赛专项
○ F、研修专项</td></tr>
<tr><td>申请金额</td><td colspan="2">20000</td><td>项目期限</td><td colspan="4">2022 年 4 月-2023 年 4 月</td></tr>
<tr><td>负责人</td><td>陈述</td><td>性别</td><td>男</td><td>民族</td><td>汉</td><td>出生年月</td><td>2002 年 8 月</td></tr>
<tr><td>学号</td><td>9203012003</td><td>联系电话</td><td colspan="5">手机:13905663645　　QQ :1628196379</td></tr>
</table>

	姓 名	学号	学院	专业班级	联系电话	项目分工
项目组成员	陈述	9203012003	理学院	统计201	13905663645	编写程序
	王艺轩	23320101	理学院	统计201	18013303671	构建模型
	王浩昱	23320102	理学院	统计201	19953170117	数据搜集与处理
	戴光裕	23220231	理学院	应化202	18718038978	提供化学专业知识
	王天声	23220201	理学院	应化202	18715577102	提供化学专业知识
指导教师	姓名	工号	学院/部门	职称	联系电话	电子邮件
	骈聪	2019057	理学院	副教授	18260080882	piancong@njau.edu.cn
负责人	无					

曾经参与科研的情况	
指导教师承担科研课题情况	1. 南京农业大学高层次人才科研启动经费 100 万元。 2. 浙江省自然科学基金委，重点项目，Z18C60003，miR+信号通路的整合与数据挖掘， 2018-01 至 2021-01，40 万元，在研，参加（第一主要参与人）。 3. 浙江省博士后，浙江省博士后科研项目择优资助，517000-X81701，基于 RNA-Seq 数据的 lncRNA 与癌症关系的研究，2018-01 至 2019-01，3 万元，结题，主 持。 4. 中国博士后，香江学者计划，188020-170257701/024， 胃癌基因组学与发展：以病人活组织培养三维类器官细胞团作整合性基因组分析、药物敏感测定、细胞生物学及动物模型研究，2017-10 至 2019-10，60 万元，结题，主持。 5. 国家自然科学基金委员会，面上项目，11571173，弱乘子（Hom-）Hopf 代数上 Galois 理论与同调理论研究，2016-01 至 2019-12，65 万，结题，参与。 6. 江苏省自然科学基金委，面上项目，BK20141358，弱 Hopf 代数上的结构和表示及同调维数研究，2014-07 至 2017-06，10 万，结题，参与。
指导教师对本项目的支持情况	积极指导学生完成此项目，并支持学生参加与本课题相关的生物信息学会议。

二、 项目内容（可加页）

（一）研究目的

随着当今时代的发展，不法分子将目光锁定在了芬太尼及其衍生物，合成第三代毒品。一方面，因芬太尼及其衍生物本身结构容易被修饰改造，不法分子会合成新型芬太尼类物质以躲避监管；另一方面，目前数据库现有的芬太尼类衍生物种类过少，难以识别。因此，为防止新型毒品走私，找出潜在的芬太尼衍生物和拓宽现有的数据库迫在眉睫。

然而，现有的多数预测方法并不能很好的生成生物分子和化学小分子，对芬太尼衍生物的预测也还缺乏一定的准确性[1]。为了更好的克服上述缺点，本课题拟使用深度学习方法，并选用SeqGAN模型对芬太尼衍生物进行预测。首先，近年来不断发展，深度生层模型在生成生物分子和化学小分子取得了一些进展，许多新的深度生成模型不断被提出。其次，我们使用了现有的芬太尼及其衍生物的数据训练深度生成模型，可以并利用训练好的深度生成模型生成分子。运用好深层生成模型，对我们后续的研究数据有着非同寻常的价值意义。

深度生成模型主要用到的是：基于SeqGAN模型（序列对抗网络）生成芬太尼类衍生物。SeqGAN就是利用生成对抗网络结合强化学习的方法来实现序列数据的生成。

SeqGAN的生成器是具有长短时记忆单元的循环递归神经网络，该模型的判别器是一个卷积神经网络。SeqGAN整体可以看成一个强化学习系统，生成器是强化学习的智能体，判别器是环境，每当智能体执行某一项动作之后，当前环境的状态发生改变，环境会给智能体奖励。

（二）研究内容

1. 了解分子生成模型，运用GAN模型，根据已有的芬太尼小分子，通过生成器和判别器进行新的芬太尼小分子的生成和判别，致力于研究出新的芬太尼分子填充到现有的芬太尼数据库中。
2. 学会使用代码，用SeqGAN模型进行预训练和最后的研究。
3. 探索深度学习的操作方法，学习相关理论知识，再设计和构建芬太尼小分子。
4. 学习python代码，了解并熟悉VAE和RNN代码，运用这两种方式进行分子生成模型的进一步探究芬太尼及芬太尼类衍生物的种类。
5. 选用合适的评估标准，对GAN模型的预测效果和性能进行分析和评估。
6. 对建立的模型进行适当的优化和改进，以提高其预测能力，拓宽模型在化学、数学等领域的应用。

（三）国内外研究现状和发展动态

近年来，深度生成模型得到了广泛应用，并在各个方面都取得了显著成果，从改变图像中的面部表情到源图像和目标图像之间的转换，生成模型在表示和生成连续域中的数据方面具有优越的性能。对于更复杂和离散的数据类型，如化学分子，开发生成模型来生成真实有效的数据受到越来越多关注。深度生成模型的发展已经产生了一系列有希望解决生成分子问题的方法[3-6]，为生成芬太尼类衍生物提供了一个有前景的新方向。

1. 基于循环神经网络的深度生成模型

由于使用简化分子线性输入规范（Simplified Molecular Input Line Entry System，SMILES）表示的分子是基于序列的，因而循环神经网络（Recurrent Neural Network，RNN）可以用于训练和生成 SMILES 字符串。2017 年，Jaques 等人[7]提出了一种序列生成结构，不仅能有效地避免生成重复的 SMILES 字符串，还能在生成文本和音乐的过程中避免生成重复样本，通过惩罚不利于分子合成的结构，增加生成有效分子的数量。同年，Olivecrona 等人[8]使用增强情景似然（AugmentedEpisodic Likelihood）和传统的策略梯度方法来调整 RNN 中分子的生成，该模型

在 ChEMBL[4]数据集上训练后生成的有效分子比例可以达到 94%。另外，该文献设计奖励函数使生成的分子避免含有硫的官能团，并生成类似于给定结构或具有特定目标活性的结构。2018 年，Popova 等人[9]使用 RNN 设计药物分子，结合强化学习（Reinforcement Learning，RL）来调整生成分子合成难易度系数、溶解性时，文献模型表现出了更好的学习 SMILES 字符串的能力。然而在强化学习训练过程中该模型容易出现“模式崩溃”（Mode Collapse），导致生成的新分子特定性质值不高。上述三篇将 RNN 模型作为深度生成模型的文献，对设计合理的奖励函数要求较高，如果奖励函数不合理将直接导致生成的分子失去实际应用价值，并且结合强化学习对分子性质进行优化时，容易出现模型崩溃，导致在强化学习训练时不能进行有效的搜索，造成生成的新分子特定性质值不高。

2. 基于变分自编码器的深度生成模型

随着分子设计抽象程度的增加，更复杂的生成模型被提出来探索化学空间。例如，变分自编码器（Variational Auto-Encoder，VAE）可以包含分子结构和分子性质之间的直接映射。VAE 的编码器和解码器的联合训练能够使用实数和压缩表示逼近非常复杂的数据分布，这对于改进化合物的搜索有着积极意义。由于潜在空间是有意义的，编码器学习将潜在空间中的向量与真实数据的属性联系起来。在编码器和解码器共同训练后，生成模型可以与推理步骤解耦，并将潜在变量作为研究领域。因此，VAE 将原始的化学空间映射到一个连续的、可微的空间，传递了原始分子的所有信息。2016 年，Bombarelli 等人提出 CVAE（Chemical-VAE）模型，使用该模型对 QM9 和 ZINC 数据集中选取的分子进行预测和重建。潜在空间不仅允许分子采样，而且允许使用在潜在空间上训练的高斯过程预测器进行插值、重建和优化。2017 年，Kusner 等人[10]提出 GVAE（Grammar-VAE）模型，该模型在 VAE模型的基础上加入上下文无关语法（Context Free Grammars）[11]，但有效分子的比例只是提升到了 31.0%。2018 年，Hanjun 等人[12]受编译器理论的启发，引入随机惰性属性（Stochastic Lazy Attributes），提出了一种基于 VAE 改进的 SD-VAE（Syntax-Direted-VAE）模型，这种方法将离线的 SDT（Syntax Directed Translation）检测转化为实时生成的指导，对解码器进行约束，最后生成的有效分子比例可以达到 43.5%。2018 年，Jin 等人[13]针对深度生成模型生成有效分子比例低的问题，提出使用分子子结构对 VAE 进行训练，训练后的模型生成的有效分子比例可以达到 100%，然而结合贝叶斯优化方法对分子特定性质进行时，生成的新分子特定性质值并不高。2020 年，Yan 等人[14]提出在重构损失项前加一个 α 系数（该系数大于 1），用来平衡 KL 散度（Kullback－Leibler Divergence）损失，缓解后验损失问题，使生成的新分子比例可以达到 90%。但这种方法存在两个问题，（1）该系数需要反复调整；（2）文献只针对一个分子数据集进行训练，没有验证模型的泛化能力。可以很清楚地发现，基于 VAE 改进的 CVAE、GVAE 和 SD-VAE 模型，生成的有效分子比例都较低，这主要是因为上述文献并没有解决VAE 存在

的“后验失效”问题，Yan 等人[14]提出通过调节一个系数来缓解“后验失效”问题，然而并没有在不同数据集上验证其泛化能力，且该系数需要反复调节。虽然 Jin 等人[13]提出使用分子子结构训练改进的 VAE 模型，生成的有效分子比例可以达到 100%，然而生成的新分子范围受限于给定的子结构，造成优化的新分子特定性质值并不高。

3. 基于生成对抗网络的深度生成模型

生成对抗网络是当前机器学习深度学习领域最热门的研究方向[3]。由于生成对抗网络一开始是处理图像生成的问题，故生成对抗网络用到序列生成上会导致生成对抗网络难以训练。为了解决这个问题，Lantao Yu 等人提出了基于生成对抗网络和强化学习的SeqGAN，该模型通过强化学习绕过采样带来的问题，从而使得生成对抗网络可以用于文本序列的生成[15]。除此之外，还有一些专门用于生成化学小分子的生成对抗网络模型被提出来。2017 年，Guimaraes 等人提出 ORGAN（Objective-ReinforcedGAN）模型，该模型结合随机策略，使生成的新分子具有多样性。但在训练过程中不稳定，生成有效分子比例只有 37.9。2018 年，Cao 等人提出 MolGAN（Molecular GAN）模型，该模型将 GAN 和 RL 结合用于训练生成分子图，但MolGAN 模型太容易出现“模式崩溃”的问题，生成有效的分子中近 90% 是重复的。

（四）创新点与项目特色

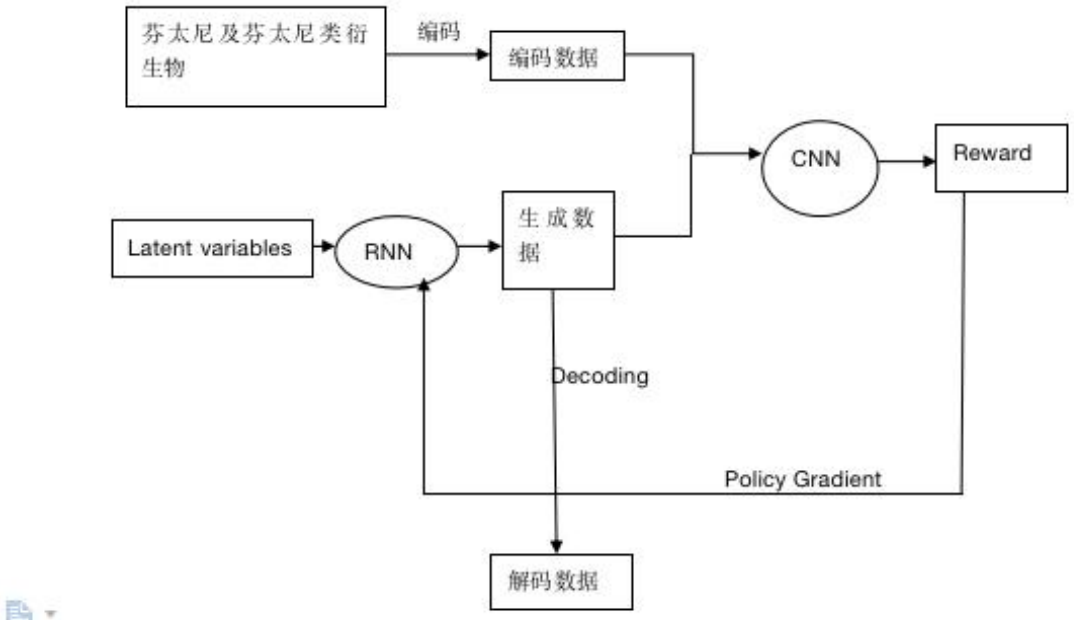
1. 研究方法方面:本课题使用了现有的芬太尼及其衍生物的数据训练了深度生成模型，并利用训练好的深度生成模型生成分子。
2. 研究对象方面:本课题选取了芬太尼作为研究对象，芬太尼及其衍生物因滥用而变成第三代毒品。因芬太尼及其衍生物本身结构就容易被修饰，不法分子会合成新型芬太尼类物质以躲避监管。而数据库现有的芬太尼类衍生物种类过少，因此研究芬太尼衍生物的生成极具现实意义。
3. 研究内容方面:本课题是基于SeqGAN模型（序列对抗网络）生成芬太尼类衍生物，根据已有的芬太尼小分子，通过生成器和判别器进行新的芬太尼小分子的生成和判别，从而研究出新的芬太尼分子。
4. 样本数据方面:选用pubchem数据库和文献中的芬太尼衍生物数据，并对数据进行去重，将这个数据集作为本课题的数据集,有助于提高模型预测的准确性。

参考文献：

[1]Xavier DA, Carlos G, Elizabeth B, J.Rafael MB, Winnie U, Aries E.A, Emily C, H.Paul B, Gary S. The METLIN small molecule dataset for machine learning-based retention timeprediction[J]. Nature Communications, 2019, 10(1).
[2]GAN
[3] G. L. Guimaraes, B. Sanchez-Lengeling, C. Outeiral, et al. Objective-reinforced generative adversarial networks (organ) for sequence generation models[J]. arXiv preprint arXiv:1705.10843, 2017
[4] R. Gómez-Bombarelli, D. Duvenaud, J. M. Hernández-Lobato, et al. Automatic chemical design using a data-driven continuous representation of molecules[J]. ACS Central Science, 2016, 4(2)
[5] N. De Cao, T. Kipf. Molgan: An implicit generative model for small molecular graphs[J]. arXiv preprint arXiv:1805.11973, 2018
[6] J. You, B. Liu, R. Ying, et al. Graph convolutional policy network for goal-directed molecular graph generation[J]. arXiv preprint arXiv:1806.02473, 2018
[7] N. Jaques, S. Gu, D. Bahdanau, et al. Sequence tutor: Conservative fine-tuning of sequence generation models with kl-control[C]. International Conference on Machine Learning, 2017, 1645-1654
[8] M. Olivecrona, T. Blaschke, O. Engkvist, et al. Molecular de-novo design through deep reinforcement learning[J]. Journal of Cheminformatics, 2017, 9(1): 48
[9] M. Popova, O. Isayev, A. Tropsha. Deep reinforcement learning for de novo drug design[J]. Science advances, 2018, 4(7): eaap7885
[10] M. J. Kusner, B. Paige, J. M. Hernández-Lobato. Grammar variational autoencoder[J]. arXiv preprint arXiv:1703.01925, 2017
[11] H. John. Introduction to automata theory, languages, and computation[J]. Acm Sigact News, 2001, 32(1): 60-65
[12] H. Dai, Y. Tian, B. Dai, et al. Syntax-directed variational autoencoder for structured data[J]. arXiv
preprint arXiv:1802.08786, 2018
[13] W. Jin, R. Barzilay, T. S. Jaakkola. Junction tree variational autoencoder for molecular graph generation[C]. International Conference on Machine Learning, 2018, 2323-2332
[14] C. Yan, S. Wang, J. Yang, et al. Re-balancing variational autoencoder loss for molecule sequence
generation[C]. Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, 2020, 1-7
[15] SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient

（五）技术路线、拟解决的问题及预期成果

1. 技术路线



运用SeqGAN模型进行芬太尼小分子的序列生成和判别。SeqGAN整体可以看成一个强化学习系统，生成器是强化学习的智能体，判别器是环境（其中生成器和判别器都是神经网络，可以进行更改和替换）。我们将Latent variables投入到生成器中生成的数据会进一步到判别器中进行判别。

为了提高SeqGAN中对抗训练效率，我们最开始将对生成器和判别器进行训练，让生成器生成的数据比较逼真，让判别器明白什么样的数据是生成数据。训练结束后，再让生成器生成一组数据作为我们的最终生成数据。

Table1：字典

Number	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Vocabulary	^	H	B	c	C	n	N	o	O
Number	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
Vocabulary	p	P	s	S	F	Q	W	I	[
Number	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**
Vocabulary	]	+	u	y	~	!	&	Z	X
Number	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**
Vocabulary	-	=	#	.	(	)	1	2	3
Numb	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**

er									
Vocabulary	4	5	6	7	8	@	/	\\	_

2. 拟解决的问题

学习VAE、RNN

下载pubchem数据库，下载几个芬太尼类衍生物的数据对其进行进一步探究

通过SeqGAN将芬太尼类衍生物进行进一步生成和判别

利用代码将生成的数字序列解码成smiles格式

3. 预期成果

深度理解生成模型SeqGAN的原理，将其应用于芬太尼的生成

发表一篇5分以上SCI论文

（六）项目研究进度安排

2022 年 1-3月:积极与导师沟通，向研究生学姐请教，进行srt初步了解学习，探索课题的具体意义和研究背景，开始学习科研项目所必备的技术，并且不断完善学习计划。

2022 年 4-5 月:寻找和搜集与课题相关的文献和实验数据，继续进行相关知识学习，了解生成模型和芬太尼及其衍生物的计算机代码，确定完善整个研究框架。

2022 年 6-9 月:小组进行分工，对实验数据进行预处理，对生成分子和芬太尼衍生物的Murcko骨架模型进行构建并开始编写代码。

2022 年 10-12 月:利用代码将生成的数字序列可以解码smiles格式，得到我们最终的生成数据。评估模型的预测效果并进行适当优化。

2023 年 1-2023 年 2 月:与导师交流讨论，分析整理实验结果，总结一年来的学习成果和经验，完成报告和论文的撰写，准备结题工作。

（七） 已有基础尚缺少的条件解决方法：

1. 已有基础：

与本项目有关的研究积累和已取得的成绩和已具备的条件：

（1）已学习数学分析、高等代数、概率论、数理统计等数学类课程，具有较好的逻辑能力和学习能力；已学习C语言程序设计和R语言，正在学习Matlab和Python,具有一定的计算机基础知识和编程能力。

（2）指导老师给予的相关指导，本课题是指导教师当前正在开展的前沿科研课题，研究思路新颖，富有科学价值，指导经验丰富。导师主要从事差异甲基化区域的算法开发，机器学习算法。在生物信息中的应用，以及非编码 RNA 与表观遗传相关性的研究，为本项目深入开展奠定良好的工作基础与学术积累。

（3）通过查阅书籍，网上浏览等途径已经对课题的研究有了一定的了解和学习。

（4）教务处、图书馆、理学院和数学系积极创造各种便利条件，积极营造立德树人、风清气正的教学科研氛围，大力支持学生创新创业训练活动。

2. 尚缺少的条件：

本项目SRT小组成员专业功底有待夯实，还需多阅读文献巩固基础知识。对课题所需的知识量还不够。缺少关于深度学习、Python 编程、SeqGAN等方面的知识与实际经验。

3. 解决方法：

充分利用互联网等途径，向指导老师寻求相关帮助，利用课余时间尽快补足知识缺口，提升自身的编程能力，早日完成项目研究。

三、 经费预算

开支科目	预算经费（元）	主要用途	阶段下达经费计划（元）	
			前半阶段	后半阶段
预算经费总额	20000.00	无	8500.00	11500.00
1. 业务费	13000.00	无	5000.00	8000.00
（1）计算、分析、测试费	2000.00	无	1000.00	1000.00
（2）能源动力费	0.00	无	0.00	0.00
（3）会议、差旅费	4000.00	国内学术交流调研	2000.00	2000.00
（4）文献检索费	1000.00	文献检索和上网费	500.00	500.00
（5）论文出版费	6000.00	论文发表的版面费	1500.00	4500.00
2. 材料费	7000.00	研究中易耗品购置	3500.00	3500.00
学校拨款				
财政拨款				

四、 项目组成员签名

五、 指导教师意见

导师（签章）： 年　月　日

六、 院系大学生创新创业训练计划专家组意见

教学负责人（签章）： 年　月　日

七、 学校大学生创新创业训练计划专家组意见

负责人（签章）： 年　月　日