腾讯 AI Lab 犀牛鸟专项研究计划申请书模版

Tencent AI Lab Rhino-Bird Focused Research Program Proposal Template

Note: You may use either the Chinese template, or the English one at the end.

（要求提交.docx 文档格式）

一、 课题负责人基本信息

姓名(中文)	许洪腾	姓名(拼音)	Xu, Hongteng
职称	副教授		
学校	中国人民大学		
院系/实验室	高瓴人工智能学院		
电子邮件	hongtengxu@ruc.edu.cn		
手机号码	13691142199		
微信帐号	Hunter313_X		
个人主页	https://sites.google.com/view/hongtengxu		
快递地址	北京市海淀区中关村大街 59 号中国人民大学高瓴人工智能学院		

二. 研究计划

腾讯伙伴	黄俊洲
申请主题	1 Machine Learning for Life Science 1.1 Machine Learning for Retrosynthesis.
研究题目	基于图最优传输理论的药物合成研究
研究背景	逆向合成是将目标分子逐步分解为一系列关键成分或小分子的过程。通过预测目标分子可能的分解步骤和相应的合成路径，逆向合成技术可以辅助设计和优化药物合成过程，有利于提高药物研发的自动化程度，提高药物生产效率，降低生产成本。 针对逆向合成问题，传统方法往往依赖基于规则的专家系统，利用启发式算法对给定的一系列关键成分进行组合，进而枚举目标分子可能的分解模式[1，2]。这类方法需要大量的人工的先验知识，往往只能对少数特定的目标分子适用，缺乏泛化能力。随着以深度学习为代表的人工智能技术的发展，越来越多的基于机器学习技术的逆向合成方法被提出[3，4]。从数据表征的角度这类方法可以分为两套技术路线：一种是利用SMILE将化学分子表示为符号序列，在通过序列神经网络，transformer等技术将逆向合成

问题转化为类似自然语言处理中的序列预测和生成问题[5，6]；另一种则是利用化学分子的图结构信息，使用图神经网络和强化学习等技术将逆向合成问题转化为有条件的图生成问题，对目标分子直接进行分解或以目标分子为输入生成关键成分[7，8]。虽然上述基于神经网络的方法取得了令人鼓舞的成果，但是它们也继承了神经网络模型的可解释性弱、过拟合风险高等问题，进而影响了它们的实际应用。近期有一些工作开始尝试将传统的符号规则策略与深度学习技术相结合[4，9，10]，并对上述问题取得了一些改进。但是，要从根本上实现具有鲁棒性和可解释性的逆向合成方法还需要从理论层面设计新的模型和算法框架。

[1] Delépine, Baudoin, et al. "RetroPath2. 0: A retrosynthesis workflow for metabolic engineers." Metabolic engineering 45 (2018): 158-170.

[2] Yuan, Shuai, et al. "Retrosynthesis of multi-component metal-organic frameworks." Nature communications 9.1 (2018): 1-11.

[3] Coley, Connor W., et al. "Computer-assisted retrosynthesis based on molecular similarity." ACS central science 3.12 (2017): 1237-1245.

[4] Segler, Marwin HS, Mike Preuss, and Mark P. Waller. "Planning chemical syntheses with deep neural networks and symbolic AI." Nature 555.7698 (2018): 604-610.

[5] Karpov, Pavel, Guillaume Godin, and Igor V. Tetko. "A transformer model for retrosynthesis." International Conference on Artificial Neural Networks. Springer, Cham, 2019.

[6] Tetko, Igor V., et al. "State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis." Nature communications 11.1 (2020): 1-11.

[7] Shi, Chence, et al. "A Graph to Graphs Framework for Retrosynthesis Prediction." arXiv preprint arXiv:2003.12725(2020).

[8] Ishida, Shoichi, et al. "Prediction and interpretable visualization of retrosynthetic reactions using graph convolutional networks." Journal of Chemical Information and Modeling 59.12 (2019): 5026-5033.

[9] Segler, Marwin HS, and Mark P. Waller. "Neural-symbolic machine learning for retrosynthesis and reaction prediction." Chemistry-A European Journal 23.25 (2017): 5966-5971.

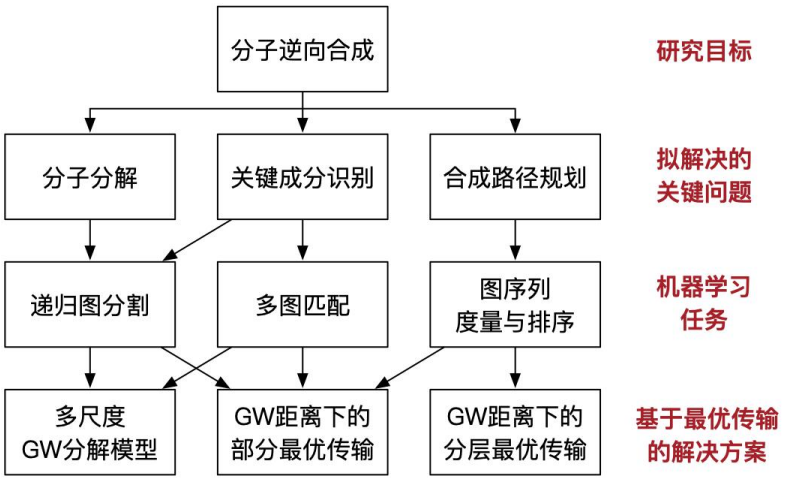
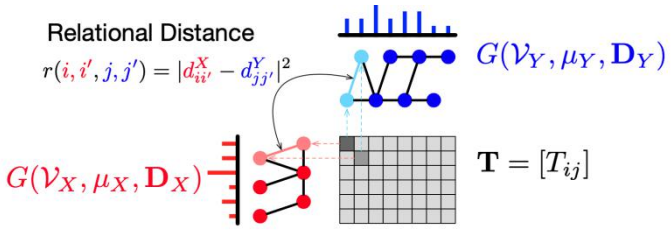
<table>
<tr><td></td><td>[10] Dai, Hanjun, et al. "Retrosynthesis prediction with conditional graph logic network." Advances in Neural Information Processing Systems. 2019.</td></tr>
<tr><td>研究目标</td><td>本项目拟针对以药物分子为代表的图结构数据，建立基于格罗莫夫-瓦瑟斯坦（Gromov-Wasserstein）距离的图最优传输理论和机器学习模型，实现具有理论支撑的图分析和生成算法，并应用于逆向合成等任务。</td></tr>
<tr><td>技术路线</td><td>本项目拟针对以药物分子为代表的图结构数据建议最优传输理论并设计相应的机器学习模型和算法。具体的研究内容包括 1）针对逆向合成问题设计多尺度格罗莫夫-瓦瑟斯坦（Gromov-Wasserstein（GW））分解模型；2）针对逆向合成问题中药物分子关键成分识别的问题设计基于非平衡GW距离的部分最优传输学习算法；3）针对不同药物分子或关键成分在合成过程中的交互和融合设计基于GW距离的分层最优传输（Hierarchical Optimal Transport）模型，实现不同合成路径之间区别的度量和合成路径排序。技术路线如图 1 所示，从上到下分别对应本项目的总目标，拟解决的关键问题，对应的机器学习任务，以及基于最优传输理论的解决方案，不同框图之间的箭头表示其继承关系。
 

图 1：技术路线图

1）多尺度GW分解模型
本项目拟利用格罗莫夫化（Gromovization）的最优传输距离来度量不同图结构数据之间的区别，并计算其节点之间的最优传输矩阵。如图 2 所示，</td></tr>
</table>

给定两个图，我们可以计算它们之间的Gromov-Wasserstein（GW）距离[1, 2]。该距离将两个图边与边之间距离看作一个随机变量并最小化其期望值。对应最小期望值的边的联合分布，即两图的边之间的最优传输矩阵，表示为两图的节点之间的最优传输矩阵的Kronecker乘积。GW距离已经被证明是图结构数据之间距离的有效度量，可以用于比较任意大小的图之间的区别[2, 3]。更重要的是，基于GW距离得到的节点之间的最优传输矩阵显性地估计了图与图之间节点的对应关系，为图匹配等任务提供了依据[3]。



The GWD is **the minimum expectation of the relational distance:**

$$d^2_{\mathrm{gw}}(G_X, G_Y) := \min_{\boldsymbol{T}\in\Pi(\boldsymbol{\mu}_X,\boldsymbol{\mu}_Y)} \mathbb{E}_{(i,i',j,j')\sim\boldsymbol{T}\otimes\boldsymbol{T}}[r(i,i',j,j')]$$
$$= \min_{\boldsymbol{T}\in\Pi(\boldsymbol{\mu}_X,\boldsymbol{\mu}_Y)} \langle \boldsymbol{D}_{XY} - 2\boldsymbol{D}_X\boldsymbol{T}\boldsymbol{D}_Y^\top,\ \boldsymbol{T}\rangle,$$

图 2：GW距离的示意图

在逆向合成问题中，本项目拟对目标分子进行逐步分解得到其一系列关键成分。其中，目标分子可以看做一种特殊的图结构数据，其节点和边都具有重要的属性信息。相应地，目标分子的逐步分解可以划归为图的递归分割问题，即将一个复杂的图在不同尺度下分割为许多子图。

针对图分割问题，本项目的前期工作设计了一种基于GW距离的图分割算法[3]，并将这种方法扩展为一个无监督的图分解模型，即GW分解模型[4]。上述工作已经在许多图结构数据的分割和分解任务中取得了良好的效果，其算法的稳定性具有理论保证。具体来说，给定一个图结构数据，通过计算其与任意“无边图”之间GW距离，我们可以得到目标图的节点与“无边图”的节点之间的最优传输矩阵。所有传输到“无边图”上同一个节点的目标图节点形成了目标图的一个子图，进而指示了目标图的分割结果。当给定一组图结构数据，GW分解模型将每个目标图表示为一组“基图”的加权GW重心图[4, 5]。通过最小化“基图”到目标图的GW距离之和，我们可以无监督地学习“基图”。

本项目拟基于上述前期工作，针对逆向合成问题设计多尺度GW分解模型。具体而言，基于GW距离的图分割有潜力实现对每个分子的成分分割，自动得到一系列小分子结构。当给定一组分子时，GW分解模型则可以帮助我们学习这些分子通用的关键成分作为“图基”。如图 3 所示，对于学习得到的“图基”，我们可以迭代使用GW图分割或GW分解技术，进一步得到更小尺度的子图或“图基”。这种多尺度的GW分解模型可以帮助我们在不同尺度下估计目标分子的关键成分 —— 尺度越小，对应的关键成分越基础，所估计的合成阶段越初级。

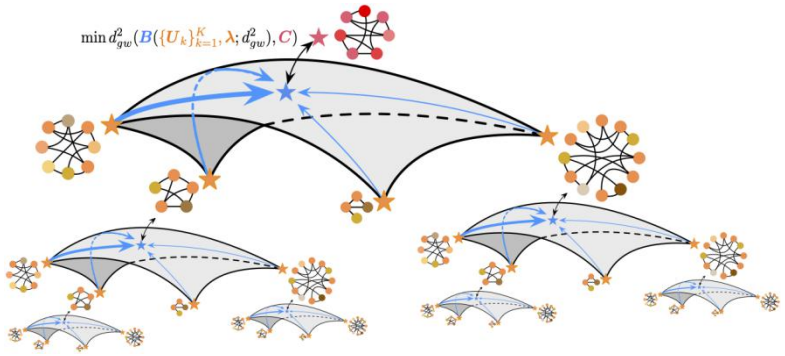


图 3：多尺度GW分解模型的示意图

2）基于GW距离的部分最优传输算法

上述多尺度GW分解模型的学习和应用需要反复计算不同大小的图（药物分子）之间的GW距离。给定两个图，传统的GW距离在计算节点之间的最优传输矩阵时会考虑每个图上所有的节点。但是针对不同化学分子计算GW距离时，我们希望通过学习它们的部分节点（即原子）之间的最优传输来识别它们共享的关键结构，如图 4 所示。从图论的角度，该问题对应子图匹配问题。

本项目拟设计一种GW距离的变种——称为“部分GW距离”——实现图与图节点之间的部分最优传输。这种新型GW距离的核心是在学习节点之间的最优传输矩阵的同时学习节点上的概率分布。我们将通过考虑节点分布的稀疏性等约束，最终得到部分节点之间的最优传输矩阵，从而实现一种新的子图匹配估计算法。当给定多个图的情况，我们可以应用上述“部分GW距离”计算“部分GW重心图”，进而实现多图的子图匹配估计算法。

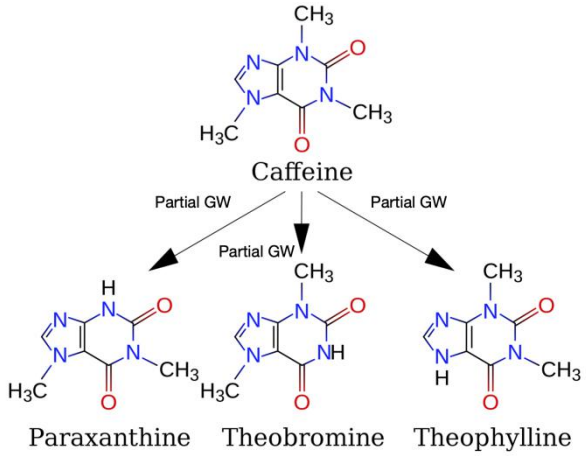


图 4：基于部分GW距离实现药物分子关键成分的匹配与识别

除了将“部分GW距离”应用于药物关键成分的识别，本项目拟从理论层面深入研究这一距离的性质，包括计算方法的鲁棒性和收敛性。在算法实现方面，本项目拟针对复杂图结构数据（如药物大分子）设计具有低复杂度的“部分GW距离”计算算法。上述研究将为基于GW距离的逆向合成方法提供理论和技术支持。

3）GW距离下的分层最优传输模型

针对目标分子的逆向合成结果，我们需要对一系列可能的合成路径进行评估和排序，选择最有可能的合成路径。如图 5 所示，每一条合成路径都可以表示为一个图序列，其节点为各个阶段的关键成分，边为合成方法。因此，合成路径的筛选问题可以建模为图序列的距离度量与排序问题。

本项目拟设计一种基于GW距离的分层最优传输模型实现上述图序列的距离度量和排序。如图 5 所示，给定两条不同的合成路径对应的图序列，我们可以计算不同序列节点之间的距离和它们边之间的距离。其中节点间的距离对应不同分子之间的GW距离，边之间的距离为合成方法对应的特征向量之间的距离加上不同分子间的GW距离。需要说明的是，这里的GW距离可以是本项目拟设计的部分GW距离。将上述基于GW设计的距离作为底层距离（Ground distance），我们可以综合考虑距离序列节点集合之间的Wasserstein距离和边集合之间的GW距离，计算上述图序列的融合GW（Fused GW）距离[6]。上述模型设计了一种分层结构，基于合成路径上不同步骤之

间的最优传输设计了不同合成路径之间的最优传输。

利用上述分层最优传输模型，我们可以计算不同合成路径之间的距离，进而构造合成路径的核矩阵（Kernel matrix），实现合成路径的聚类分析、排序、筛选等任务。

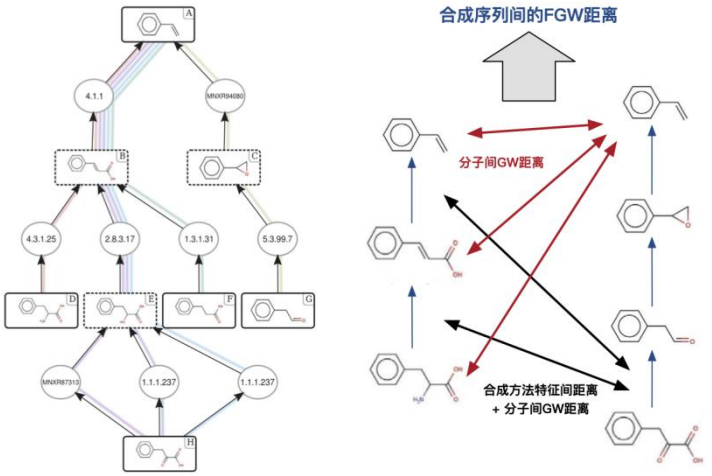


图 5：合成路径示意图与GW分层最优传输模型，左图摘自文献[7]

4）本项目的潜在优势

（a）基于GW距离的性质和对应的**图最优传输理论**，本项目将逆向合成问题化归为一系列图分析和图生成问题，其方法基于GW距离和对应的图最优传输理论，具有坚实的理论支撑。

（b）本项目所提出的方法可以显性地计算分子与其关键成分之间，分子之间，乃至合成路径之间的最优传输距离。基于最优传输矩阵，这些方法得到的结果相比现有方法具有更好的**可解释性**。

（c）本项目拟提出的图最优传输理论和技术具有**普适性**，除分子逆向合成问题以外，还可以应用于药物的毒理学分析（图分类、图聚类问题）等其他任务。

（d）本项目拟取得的研究成果与现有的深度学习方法有较高的**兼容性**。上述GW距离及其变种可以作为图神经网络等深度学习模型的目标函数或正则项，辅助其学习并得到更好的结果。

项目所需资源

(1) 数据资源与软件工具

充分利用开源的化学反应过程与逆过程的数据库,如 RetroTransformDB [8]、QCArchive [9]等，以及开源的分子数据分析软件，如 DeepChem, ChemoPy, RDKit 等等。同时对 ChemSub [10]等在线开源数据集进行自动抓取或人工收集。

(2) 合作关系

Lawrence Carin，杜克大学电子与计算机工程系、教授、副校长（近期拟入职阿普杜拉国王科技大学担任教务长）。

查宏远，香港中文大学深圳校区数据学院，教授

Ricardo Henao，杜克大学生物信息与生物统计系，助理教授。

Xia Ning，俄亥俄州立大学计算机科学系与生物信息系，副教授。

David Page，杜克大学生物信息与生物统计系，教授，系主任。

Le Song，佐治亚理工学院计算科学与技术系，副教授。

以上研究者在计算机辅助的生物医药研究等领域取得了许多前沿成果，并与项目组成员在机器学习及应用方向有长期合作关系，可以针对本项目在数据共享和技术创新方面展开深度合作。

(3) 计算资源

GPU 服务器等相关机器资源。

参考文献

[1] Mémoli, Facundo. "Gromov–Wasserstein distances and the metric approach to object matching." Foundations of computational mathematics 11.4 (2011): 417-487.

[2] Chowdhury, Samir, and Facundo Mémoli. "The Gromov-Wasserstein distance between networks and stable network invariants." Information and Inference: A Journal of the IMA8.4 (2019): 757-787.

[3] **Xu, Hongteng,** Dixin Luo, and Lawrence Carin. "Scalable Gromov-Wasserstein learning for graph partitioning and matching." Advances in neural information processing systems. 2019.

<table>
<tr><td></td><td>[4] Xu, Hongteng - "Gromov-Wasserstein Factorization Models for Graph Clustering", AAAI Conference on Artificial Intelligence, 2020.

[5] Peyré, Gabriel, Marco Cuturi, and Justin Solomon. "Gromov-Wasserstein averaging of kernel and distance matrices." International Conference on Machine Learning. 2016.

[6] Vayer, Titouan, et al. "Fused Gromov-Wasserstein distance for structured objects." Algorithms 13.9 (2020): 212.

[7] Delépine, Baudoin, et al. "RetroPath2. 0: A retrosynthesis workflow for metabolic engineers." Metabolic engineering 45 (2018): 158-170.

[8] Avramova, Svetlana, Nikolay Kochev, and Plamen Angelov. "RetroTransformDB: A Dataset of Generic Transforms for Retrosynthetic Analysis." Data 3.2 (2018): 14.

[9] https://qcarchivetutorials.readthedocs.io/en/latest/cookbook/overview.html

[10] http://chemsub.online.fr</td></tr>
<tr><td>计划进度</td><td>关键时间节点及该阶段产出。
2021.4-2021.6
对药物逆向合成问题的相关工作进行全面调研；基于 GW 距离及其变种，设计药物分子的分解模型与无监督学习算法。
2021.7-2021.9
以所提出 GW 分解模型为基础，研究药物分子的递归分解方法，设计相应的多尺度 GW 分解模型与基于 GW 距离的部分最优传输学习算法，完成中期报告。
2021.10-2021.12
设计基于 GW 的分层最优传输模型和学习算法。针对以所提出的模型和算法，研究其理论性质并在药物逆向合成问题上进行实验验证。结合神经网络技术，将所提出的方法扩展到药物组合分析，受限条件下的合成等问题。
2022.1-2022.4
总结所提出模型和算法，建立代码库和针对药物逆向合成问题的系统原型，准备结题答辩。</td></tr>
<tr><td>预期产出</td><td>技术储备：未来一年基于PyTorch建立GW最优传输模型和算法的代码库与系</td></tr>
</table>

	统原型。针对相应的模型和算法申请 1 项专利。 学术影响：未来一年在机器学习领域的顶级会议或期刊投稿 2-3 篇论文，例如NeurIPS2021，ICLR2022，ICML2022，AAAI2022，IEEE Trans，JMLR等。 人才培养：计划安排 1 名学生利用假期进行实习，进行联合培养。

三. 项目组成员及相关研究背景

项目组成员（请勿填写不具体从事该课题的实验室其它人员）

姓名	职称（老师） 年级（学生）	手机	邮箱
许洪腾	**副教授**	**13691142199**	**hongtengxu@ruc.edu.cn**

项目组成员**相关研究背景**

（部分项目组成员**近三年**与该项目密切相关的经验和成果，比如发表的文章等）

许洪腾，中国人民大学，高瓴人工智能学院，副教授。主要研究兴趣包括机器学习理论及其应用。自 2011 年起已著有近 60 篇国际顶尖期刊和会议论文，一项国际专利，其中以第一作者身份在机器学习领域顶级会议和期刊发表论文 20 篇，包括ICML，NeurIPS，AAAI，IJCAI，CVPR，ICCV，IEEE Trans等。于 2019 年提出的格罗莫夫-瓦瑟斯坦学习方法在图分析领域建立了新的基于最优传输理论的机器学习框架，得到了国内外同行的广泛关注。目前担任IEEE TNNLS期刊客座编辑，多个国际期刊的审稿人和会议的技术委员。近三年与最优传输理论以及图分析模型有关的研究成果包括：

[1] Yujia Xie, Yixiu Mao, Simiao Zuo, **Hongteng Xu**, Xiaojing Ye, Tuo Zhao, Hongyuan Zha - “A Hypergradient Approach to Robust Regression without Correspondence”. International Conference on Learning Representations (**ICLR**), 2021.

[2] **Hongteng Xu**, Dixin Luo, Lawrence Carin, Hongyuan Zha - "Learning Graphons via Structured Gromov-Wasserstein Barycenters", AAAI Conference on Artificial Intelligence (**AAAI**), 2021.

[3] Wenlin Wang, **Hongteng Xu**, Guoying Wang, Wenqi Wang, Lawrence Carin - “Zero-Shot Recognition via Optimal Transport”, IEEE Winter Conference on Applications of Computer Vision (**WACV**), 2021.

[4] **Hongteng Xu**, Dixin Luo, Ricardo Henao, Svati Shah, Lawrence Carin - "Learning Autoencoders with Relational Regularization", The International Conference on Machine Learning (**ICML**), 2020.

[5] Xuan Zhang, Shaofei Qin, Yi Xu, and **Hongteng Xu** - "Quaternion Product Units for Deep Learning on 3D Rotation Groups", IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), 2020.

[6] **Hongteng Xu** - "Gromov-Wasserstein Factorization Models for Graph Clustering", AAAI Conference on Artificial Intelligence (**AAAI**), 2020.

[7] Wenlin Wang, **Hongteng Xu**, Zhe Gan, et al. - "Graph-Driven Generative Models for Heterogeneous Multi-Task Learning", AAAI Conference on Artificial Intelligence (**AAAI**), 2020.

[8] **Hongteng Xu**, Dixin Luo, Lawrence Carin - "Scalable Gromov-Wasserstein Learning for Graph Partitioning and Matching", The Conference on Neural Information and Processing System (**NeurIPS**), 2019.

[9] **Hongteng Xu**, Dixin Luo, Hongyuan Zha, Lawrence Carin - "Gromov-Wasserstein Learning for Graph Matching and Node Embedding", The International Conference on Machine Learning (**ICML**), 2019.

[10] **Hongteng Xu**, Wenlin Wang, Wei Liu, Lawrence Carin - "Distilled Wasserstein Learning for Word Embedding and Topic Modeling", The Conference on Neural Information and Processing System (**NeurIPS**), 2018.

[11] Xuanyu Zhu, Yi Xu, **Hongteng Xu**, Changjian Chen - "Quaternion Convolutional Neural Networks", The European Conference on Computer Vision (**ECCV**), 2018.

[12] Shuai Xiao, **Hongteng Xu**, et al. - "Learning Conditional Generative Models for Temporal Point Processes," The Thirty-Second AAAI Conference on Artificial Intelligence (**AAAI**), 2018.

[13] **Hongteng Xu**, Licheng Yu, Mark Davenport, Hongyuan Zha - "A Unified Framework for Manifold Landmarking", IEEE Transactions on Signal Processing (**TSP**), 2018.

[14] **Hongteng Xu**, Lawrence Carin, Hongyuan Zha - “Learning Registered Point Processes from Idiosyncratic Observations,” The International Conference on Machine Learning (**ICML**), 2018.

[15] Xu Chen, Yongfeng Zhang, **Hongteng Xu**, Zheng Qin, Hongyuan Zha - “Adversarial Distillation for Efficient Recommendation with External Knowledge”, ACM Transactions on Information Systems (**TOIS**), 2018.

四. 附录

若有其他需要说明的情况，请以附录形式提供。