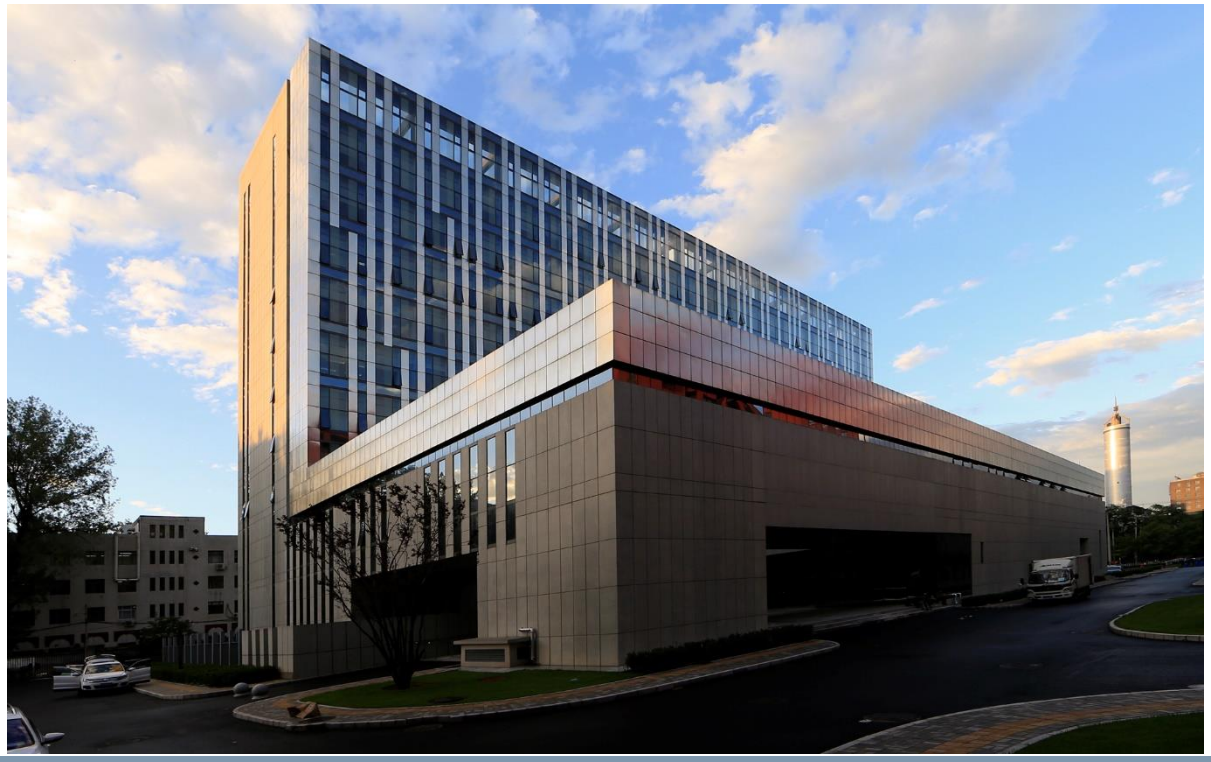
勤 数
笃 系
求 天
真 地

# 中国科学院数学与系统科学研究院

Academy of Mathematics and Systems Science
Chinese Academy of Sciences



# 人工智能与生物信息学研讨会

# BioAI'2021

## 2021年4月24日， 中国北京

组织：**陆汝钤**（中国科学院数学与系统科学研究院）**张世华**（中国科学院数学与系统科学研究院）

**周水庚**（复旦大学） **张松懋**（中国科学院数学与系统科学研究院）

地点：**中国科学院数学与系统科学研究院南楼二层 N204 会议室**

北京海淀区中关村东路（四环保福寺桥南恒兴大厦南）

线上参会：**Zoom会议号 661 0110 9127 密码 210424**

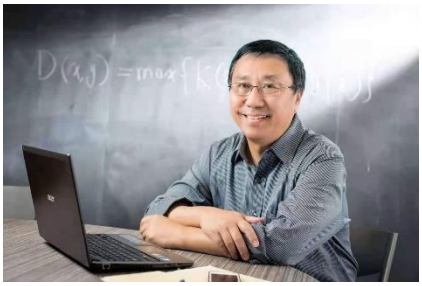
线上参会链接：https://zoom.com.cn/j/66101109127?pwd=UTBVSnBWOWdOdOMxMHBvb3B6VnBIdz09

# 研讨会程序安排

| | |
|---|---|
| 8:00-8:10 | **开幕式** |
| 8:10-11:50 上午场学术报告（共5个，主持人：张世华、张松懋） | |
| 8:10-8:50 | 人工智能赋能个体化癌症免疫治疗<br>**李明**（加拿大滑铁卢大学） |
| 8:50-9:30 | 人工智能赋能合成生物学创新发展<br>**汪小我**（清华大学） |
| 9:30-9:50 | 茶歇 |
| 9:50-10:30 | 单细胞转录组数据的分析方法介绍<br>**王飞**（复旦大学） |
| 10:30-11:10 | 大数据时代的生物信息学研究范式--以蛋白质结构预测为例<br>**卜东波**（中国科学院计算技术研究所） |
| 11:10-11:50 | 基于单细胞数据的细胞类型发现方法<br>**江瑞**（清华大学） |
| 12:00-14:00 | 午餐（物科餐厅四楼） |
| 14:00-17:30 下午场学术报告（共5个，主持人：周水庚） | |
| 14:00-14:40 | 脑效应连接网络学习研究进展<br>**冀俊忠**（北京工业大学） |
| 14:40-15:20 | 机器学习在合成致死抗癌药物靶点发现的应用<br>**郑杰**（上海科技大学 ） |
| 15:20-16:00 | Recent Advances in Machine Learning based Large-scale Protein Function Prediction<br>**朱山风**（复旦大学） |
| 16:00-16:10 | 茶歇 |
| 16:10-16:50 | Structure-Based Generative Models for *de novo* Drug Design<br>**裴剑锋**（北京大学） |
| 16:50-17:30 | 深度学习的数学理解<br>**张世华**（中国科学院数学与系统科学研究院） |

# 报告人介绍与报告摘要

## 李明 教授（加拿大滑铁卢大学）



**报告题目：**人工智能赋能个体化癌症免疫治疗

**报告摘要：**发现新抗原是个体化癌症免疫治疗的关键步骤。每个病人每个癌症每个阶段都有独特的新抗原。要实现真的癌症个体化免疫治疗，我们必须用人工智能替代湿实验室流程。

**报告人简介：**李明，加拿大皇家学会院士，ACM，IEEE，ISCB Fellow，2010年获得加拿大顶级国家科学奖Killam Prize，加拿大滑铁卢大学教授。研究Kolmogorov复杂性和计算生物学的世界权威专家。

# 报告人介绍与报告摘要

## 汪小我 教授（清华大学）

**报告题目：**人工智能赋能合成生物学创新发展

**报告摘要：**基因的编辑与合成等生物技术的出现，使得我们有可能按照功能需求从最底层设计和构建人工生物系统。如何开发与之匹配的人工智能方法，理性设计出满足多样化场景下功能需求的高性能生物功能模块，提高合成生物系统的构建效率，是人工智能与生物学交叉研究领域亟待解决的关键问题。近来，人工智能技术在全新基因元件和蛋白质结构设计方面展现出巨大潜力，有望大大扩展人工改造生命体的应用场景，推动生物制造、分子育种、基因治疗等领域的发展变革。

**报告人简介：**汪小我，博士，清华大学自动化系长聘教授。于2003年和2008年在清华大学自动化系获工学学士学位和工学博士学位，曾赴美国冷泉港实验室和加州大学伯克利分校访问学习，2008年起在清华大学任教至今。主要研究方向为模式识别与机器学习、生物信息学。担任中国生物工程学会青年工作委员会主任、中国人工智能学会生物信息学与人工生命专委会副主任、中国计算机学会生物信息学专委会常委等。曾获全国优秀博士学位论文奖、中国自动化学会青年科学家奖，并获得国家自然科学基金优秀青年基金、教育部新世纪优秀人才计划等支持。

# 报告人介绍与报告摘要

## 王飞 副教授（复旦大学）

**报告题目：**单细胞转录组数据的分析方法介绍

**报告摘要：**近年来，单细胞测序技术迅猛发展。它保留了细胞间的高度异质性，为人们研究细胞的分化、疾病的发生发展提供了更加精准的视角。其中，单细胞转录组测定单个细胞中的基因表达丰度，是目前单细胞测序技术中相对最成熟和应用最广泛的技术，成千上万、甚至百万细胞级别的数据不断涌现。单细胞数据的特点是：(1)噪音大，稀疏性高，难以区分技术偏差和生物学差异；(2)数据分布复杂，线性关系和广义线性模型难以准确捕捉数据的本质特征；(3)数据异质性高，不同细胞类型在数据中的出现频率可能相差一个数量级。本报告将介绍和评述单细胞转录组数据分析的流程，着重介绍深度学习的相关应用。

**报告人简介：**王飞，复旦大学计算机科学技术学院副教授、上海市智能信息处理重点实验室副主任。主要研究方向是人工智能、机器学习、生物信息学和系统生物学。曾主持国家自然基金面上项目四项，国家科技部项目二项。目前是计算机学会生物信息学专委会委员。

# 报告人介绍与报告摘要

## 卜东波 教授（中国科学院计算技术研究所）

**报告题目：** 大数据时代的生物信息学研究范式—以蛋白质结构预测为例

**报告摘要：** 基因组、蛋白质组等生物学数据是典型的大数据；大数据时代的生物信息学研究范式似乎发生了显著的改变：从传统“Insight->理性建模->实验结果分析”范式到“Insight->NN->实验结果分析->理解NN”范式。这种转变在蛋白质结构预测领域表现得尤为显著：从传统的基于统计模型的残基间预测算法，到纯粹基于NN的残基间距离预测。本次报告将以蛋白质结构预测为例，讲述一些思考和困惑。

**报告人简介：** 2000年于中科院计算所获得博士学位，2006-2008年与加拿大滑铁卢大学访问；研究兴趣包括生物信息学（蛋白质结构预测、糖结构鉴定）、计算机算法，在Nature Communications, NAR, AC, GUT, JPR, Bioinformatics, ISMB, RECOMB等期刊和会议发表论文多篇。研制了“用人工智能技术辅助算法设计”的AIA系统，在经典排课问题上实现了变“凭灵感设计算法”为“从数据学习出算法”；开发了“端到端”神经网络架构CopulaNet预测残基间距离，开发了蛋白质空间结构预测软件ProFOLD，性能超过AlphaFold，改进版有望达到AlphaFold2的水平。

# 报告人介绍与报告摘要

## 江瑞 副教授（清华大学）

**报告题目：**基于单细胞数据的细胞类型发现方法

**报告摘要：**细胞类型辨识是单细胞数据分析的基础，传统基于单细胞数据的研究思路是在数据降维的基础上，通过聚类确定细胞簇，再利用其基因表达谱进行细胞类型注释。近年来单细胞测序技术迅速发展，已经从最初的仅测定基因组和转录组走向更深入的表观遗传组、三维基因组、空间转录组等，如何利用这些数据来促进罕见细胞类型的发现、进行细胞功能建模，是生物信息学研究中亟待解决的关键问题。本报告将在综述已有细胞类型发现工作的基础上，介绍我们建立的一系列以深度学习为特色，以融合单细胞染色质开放性与基因表达数据为特征的细胞类型发现方法，以及后续开展的细胞类型特异基因调控网络构建、细胞周期时序分析等工作。

**报告人简介：**江瑞，2002年毕业于清华大学自动化系获工学博士学位，后在南加州大学进行博士后研究，2007年至今任清华大学自动化系副教授。从事生物信息学研究15年，致力于运用人工智能与机器学习方法研究复杂疾病的调控模式与功能建模，主要研究领域包括单细胞数据分析与细胞功能建模，基因调控模式识别，致病遗传因素分析等。

# 报告人介绍与报告摘要

## 冀俊忠 教授（北京工业大学）

**报告题目：**脑效应连接网络学习研究进展

**报告摘要：**人脑效应连接网络刻画了脑区间神经活动的因果效应。从功能磁共振成像数据中学习脑效应连接网络是人脑连接组研究中一项重要的研究课题。报告将对脑效应连接网络学习的主要流程、研究现状及分类体系、代表性方法及其性能对比进行阐述，并结合课题组完成的部分工作介绍基于蚁群优化、基于时序评分和基于递归生成对抗网络的脑效应连接学习方法。

**报告人简介：**冀俊忠，北京工业大学信息学部计算机学院教授，博士生导师。现任北京人工智能研究院副院长，多媒体与智能软件技术北京市重点实验室副主任，中国人工智能学会理事、智慧医疗专委会常务委员、不确定性人工智能专委会委员，中国计算机学会人工智能与模式识别专委会委员。主要从事机器学习、生物信息挖掘、脑影像数据智能分析等方向的研究，主持国家自然科学基金、教育部博士点基金、北京市自然科学基金等多项国家、省部级项目，在TKDE、TIP、TNNLS、PR、INF、TCBB、JBHI、AAAI、自动化学报等国际、国内权威期刊或会议上发表论文100多篇。

# 报告人介绍与报告摘要

## 郑杰 副教授（上海科技大学）

**报告题目：**机器学习在合成致死抗癌药物靶点发现的应用

**报告摘要：**合成致死（Synthetic lethality，简称SL）是基因之间的相互作用的一种关系，即同时扰动两个基因会导致细胞死亡或活力下降，而扰动任何其中一个基因都不会致命。SL反映了癌细胞和正常细胞的生物学内源性差异，因此抑制具有癌症特异性突变基因的SL伙伴基因可以选择性地杀死癌细胞而不影响正常细胞的生存。因此，SL是一个有希望发现抗癌药物靶点的金矿。湿实验筛选SL的方法受到高成本、批量效应和脱靶等问题的困扰。目前预测SL的计算方法包括基因敲除模拟、基于知识的数据挖掘和机器学习方法。

我的报告将主要包括两个部分：(1)收集了SL相关的数据和知识的SynLethDB数据库；(2)预测SL的机器学习方法。首先，我们开发了世界上第一个关于SL的系统全面的数据库SynLethDB，对于生物医药研究和制药工业是有用的资源。该数据库包含了从人类和4个模式动物的生化检测筛选、计算预测和文本挖掘结果中收集到的SL基因对。对于每一对SL基因，我们通过整合来自不同证据来源的得分计算一个综合得分。我们还开发了一个统计分析模块，基于1000多个癌细胞系的大规模基因组数据、基因表达谱和药敏谱，来评估基于SL的药物治疗对癌细胞的用药能力和敏感性。最近，我们补充了更多信息，包括用CRISPR 基因编辑技术筛选的SL，以及一个关于SL和药物的知识图谱，SynLethKG。在第二部分，我将首先对预测SL的机器学习方法做一个简要的综述，然后将介绍我们最近提出的一个基于图神经网络的KG4SL模型，它将知识图谱的消息传递机制融入到SL预测中。前面提到的SynLethKG知识图谱被整合到SL数据，以捕获SLs之间的关系并规避人工特征工程。在真实数据上的实验比较中，KG4SL模型优于所有最先进的基准模型，这表明将知识图引入GNN对SL预测有显著影响。关于KG4SL的工作已作为常规论文被生物信息学领域的顶级国际会议ISMB/ECCB 2021接收。

**报告人简介：**郑杰博士目前在上海科技大学信息科学与技术学院担任副教授、博导、研究员，并任智能医学信息研究中心主任。 他分别在浙江大学和加州大学河滨分校获得计算机科学学士和博士学位，并在美国国立卫生研究院（NIH）下属国家生物技术信息中心（NCBI）从事博士后研究。回国之前，他曾在新加坡南洋理工大学计算机科学与工程学院担任助理教授，并担任新加坡科技研究局（A*STAR）基因组研究所（Genome Institute of Singapore）客座研究员。郑杰博士长期致力于研发生物信息学算法、软件、模型等信息技术，来促进生物医学的发展。目前，郑博士在研究知识和数据混合驱动的人工智能和数据科学技术，并将其应用于AI制药、生物信息学数据分析自动化、疾病诊疗智能辅助系统等。

# 报告人介绍与报告摘要

## 朱山风 副教授（复旦大学）

**报告题目：Recent Advances in Machine Learning Based Large-scale Protein Function Prediction**

Abstract: Proteins are building blocks of life, playing many crucial roles within organisms, such as catalysing chemical reactions, co ordinating signal pathway and providing structural support to cel ls. Automated function prediction (AFP) of proteins is thus of gre at significance in biology.

AFP can be regarded as a problem of the large-scale multi-label classification where a p rotein can be associated with multiple gene ontology terms as its labels. To boost the dev elopment of effective and efficient AFP, Critical Assessment of Functional Annotation (CA FA) has been held four times to date: CAFA1 in 2010–2011, CAFA2 in 2013–2014, CAFA3 in 2015–2016 and CAFA4 in 2019–2020 (under evaluation). In this talk, I will introduce the state-of-the-art methods in large-scale AFP, as well as our recent progress in this topic, such as GOLabeler, NetGO and DeepGraphGO.

**报告人简介：**朱山风，复旦大学副教授，博士生导师。香港城市大学博士，日本京都大学博士后，美国伊利诺伊大学香槟分校访问学者，日本京都大学访问副教授。UniProt 国际科学顾问委员会委员，中国计算机学会生物信息专业委员会创始委员，中国人工智能学会生物信息与人工生命专业委员会创始委员、中国中文信息处理学会医疗健康与生物信息处理专业委员会创始委员，中国细胞生物学会生物信息与系统生物学分会理事，中国运筹学会计算系统生物学分会理事。主持或完成四项国家自然科学基金项目，以及多个国内外企业研发项目。主要研究方向为人工智能与生物医学大数据挖掘，特别是生物医学文本挖掘、蛋白功能预测、宏基因组、药物发现、免疫信息学等。相关论文在生物信息、人工智能、数据挖掘等顶级国际会议和期刊发表，如 NeurIPS, KDD, IJCAI, ISMB,, Nucleic Acids Research等。2014年-2020 年参加 BioASQ 大规模生物医学文本自动标注国际竞赛中取得六次第一名的好成绩。2017 年参加 CAFA 大规模蛋白功能自动标注国际竞赛，在全世界50多个实验室中获得第一名。

**Reference**

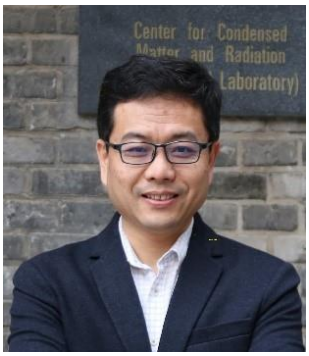
1. You et al., NetGO: improving large-scale protein function prediction with mas sive network information. Nucleic Acids Research 47(W1): W379-W387 (2019)

2. You et al., GOLabeler: improving sequence-based large-scale protein function p rediction by learning to rank. Bioinformatics 34:2465 - 2473 (2018)

3. You et al., DeepGraphGO: graph neural network for large-scale, multispecies pr otein function prediction. ISMB2021, To appear (2021)

# 报告人介绍与报告摘要

## 裴剑锋 研究员（北京大学）




**报告题目：** Structure-Based Generative Models for *de novo* Drug Design

Abstract: Recently, deep generative models for molecular graphs are gaining more and more attention in the field of de novo drug design. A variety of models have been developed to generate topological structures of drug-like molecules, but explorations in generating three-dimensional structures are still limited. Existing methods have either focused on low molecular weight compounds without considering drug-likeness or generate 3D structures indirectly using atom density maps. In this work, we introduce Ligand Neural Network (L-Net), a novel graph generative model for designing drug-like molecules with high-quality 3D structures. L-Net directly outputs the topological and 3D structure of molecules (including hydrogen atoms), without the need for additional atom placement or bond order inference algorithm. The architecture of L-Net is specifically optimized for drug-like molecules, and a set of metrics is assembled to comprehensively evaluate its performance. The results show that L-Net is capable of generating chemically correct, conformationally valid molecules with high drug-likeness. Finally, to demonstrate its potential in structure-based molecular design, we combine L-Net with MCTS and test its ability to generate potential inhibitors targeting ABL1 kinase.

**报告人简介：**裴剑锋博士，北京大学前沿交叉学科研究院特聘研究员，博士生导师。2014年起在国内率先开展人工智能药物设计研究，取得一系列研究成果，在JACS、PNAS、Nature 等国际重要学术刊物上发表论文60多篇，申请获得专利6项，软件著作权8项。主持和承担863计划、重大新药创制国家科技重大专项、基金委重点项目等国家科研项目多项。曾获中国药学会施维雅青年药物化学奖，中国化学会青年计算化学奖和药明康德生命化学研究奖。

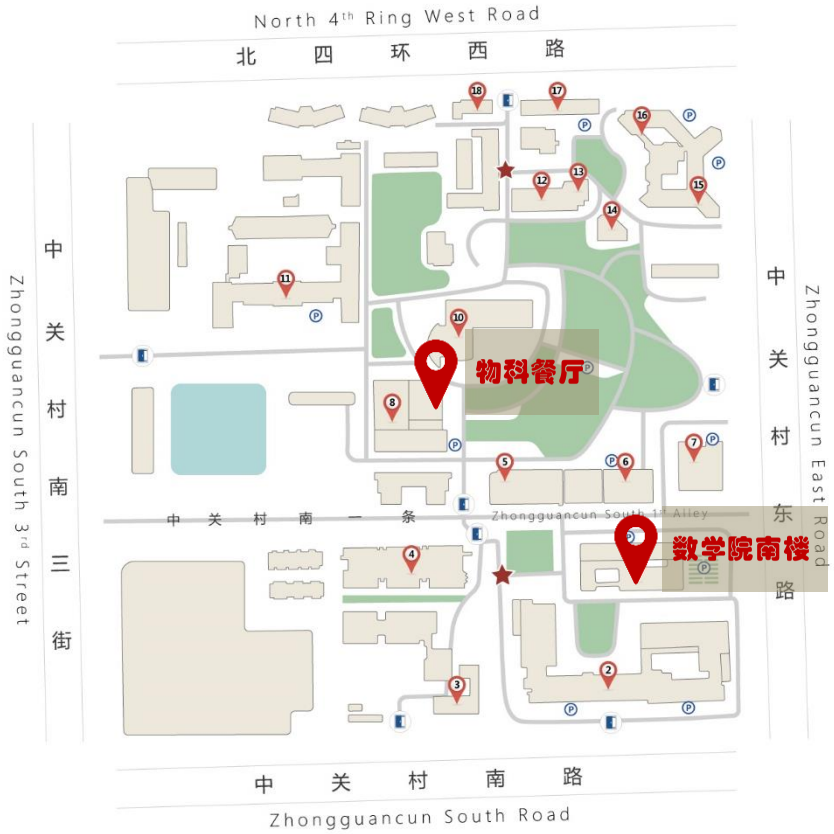
# 报告人介绍与报告摘要

## 张世华 研究员（中国科学院数学与系统科学研究院）

**报告题目：**深度学习的数学理解

**报告摘要：**深度学习特别是深度神经网络作为一种黑箱模型，是通过大量计算实验得到的，其数学原理逐渐引起研究者的广泛关注。本报告将从不同的角度介绍深度学习的数学理解与认识，特别介绍两种初步尝试。第一，从多层卷积稀疏编码模型的编码初始化和字典矩阵设计的角度，分别建立残差神经网络和多尺度密集连接网络与多层卷积稀疏编码模型的等价联系。第二，提出深度学习是在Wasserstein空间学习测地曲线的理论。在维度不变的情况下，刻画深度神经网络所学习到的映射近似最优传输映射，即数据点的表示在模型内部近似沿着直线传输，进而解释为什么残差网络相比于普通神经网络具有更好的优化和泛化能力。

**报告人简介：**张世华，中国科学院数学与系统科学研究院研究员、中国科学院随机复杂结构与数据科学重点实验室副主任、中国科学院大学岗位教授。主要从事生物信息计算、机器智能与优化，主要成果发表在Cell、Advanced Science、National Science Review、Nature Communications、Nucleic Acids Research、Bioinformatics、IEEE TPAMI、IEEE TKDE、IEEE TNNLS等杂志。目前担任BMC Genomics等杂志编委。曾荣获中国青年科技奖、国家自然科学基金优秀青年基金、中组部万人计划青年拔尖人才、中国科学院卢嘉锡青年人才奖、全国百篇优秀博士论文奖等。




大门 停车 您的位置

北

| | |
|---|---|
| 1：数学院南楼 | AMSS South Building |
| 2：物理研究所——M楼 | M Building IOP |
| 3：中国科学报社 | China Science Daily |
| 4：中国科学院大学中关村校区 | Zhongguancun Campus UCAS |
| 5：物理研究所——研究生楼 | Graduate Building IOP |
| 6：理论物理研究所新楼 | New Building ITP |
| 7：恒兴大厦 | Hengxing Mansion |
| 8：物科宾馆 | Wuke Hotel |
| 9：物科餐厅 | Wuke Dining Hall |
| 10：物理研究所——D楼 | D Building IOP |
| 11：物理研究所——A楼 | A Building IOP |
| 12：科学与工程计算国家重点实验室（科技综合楼） | LSEC ( S/T Complex Building ) |
| 13：晨兴数学中心 | Morningside Center of Mathematics |
| 14：基础科学园区餐厅 | Basic Sciences Park Dining Hall |
| 15：理论物理研究所主楼 | Main Building ITP |
| 16：数学院思源楼 | AMSS Siyuan Building |
| 17：自然科学史所 | IHNS |
| 18：科苑宾馆 | Keyuan Hotel |

会议地点：**数学院南楼二层 N204 会议室**

午餐地点：**物科餐厅四楼**