

代谢组学最新研究进展

刘俊辉

(西安交通大学 生命科学与技术学院, 710049, 西安)

摘要: 代谢组学 (metabonomics/metabolomics) 是系统生物学的分支学科之一, 也是离表型最近的组学, 主要通过通过对机体内的代谢物进行全面的定性定量分析, 阐述机体处于正常生命状态及内外环境变化后代谢过程的动态反应规律。近年来, 代谢组学在药物研发、疾病研究、植物和微生物研究等多个方面的应用取得了长足进展。其分析技术、方法上的研究突破也层出不穷。而在人工智能迅速发展的时代, 深度学习 (Deep Learning) 也在代谢组学研究中有了一席之地。

关键词: 代谢组学, 空间代谢组学, 单细胞代谢组学, 代谢流组学, 深度学习

中图分类号: Q503

一、代谢组学简介

1. 代谢组学的定义和特点

代谢组学 (metabonomics/metabolomics) 是系统生物学的分支学科之一, 也是离表型最近的组学, 主要通过通过对机体内的代谢物进行全面的定性定量分析, 阐述机体处于正常生命状态及内外环境变化后代谢过程的动态反应规律。

与以往各种组学技术相比, 代谢组学技术有其特有的优势。首先, 代谢组学能够对各类代谢产物实现高通量分析测定; 其次, 上游基因、蛋白层面的变化能在下游诸多代谢产物中得到放大, 从而使得这些变化更容易被观察到; 再者, 代谢产物的整体变化可以直接反应机体的状态。因此, 代谢组学研究在多个领域都体现出良好的应用前景。

2. 代谢组学方法及技术平台概述

完整的代谢组学分析流程包括样品的采集和预处理、数据的采集和数据的分析及解释。生物样品(如尿液、血液、组织、细胞和培养液等)采集后需进行生物反应灭活、预处理, 然后运用核磁共振、质谱或色谱等技术检测其中代谢物的种类、含量, 得到代谢谱 (metabolic profiling) 或代谢指纹 (metabolic fingerprint), 而后使用多变量数据分析方法对获得的多维数据进行降维和信息挖掘, 找寻关键代谢产物, 并研究相关的代谢途径和变化规律, 以阐述生物体对相应刺激的响应机制、发现生物标志物。

广义的代谢组学 (即非靶向代谢组学) 力求分析生物体系(如体液和细胞)中的所有代谢产物, 整个分析过程应尽可能地保留和反映总的代谢产物信息。但是, 目前代谢组学的最终目标还是不可完成的任务, 因为还没有

发展出一种代谢组学技术可以涵盖所有的代谢物而不受分子大小和性质的限制。

代谢组学技术平台主要由样品的采集和预处理技术、样品分析技术和数据处理技术三个部分组成。一个通用的、理想的代谢组学分析平台应具有以下几个特征: (1) 能够检出生物样本中尽可能多的代谢物; (2) 对检出的代谢物能够定量和定性; (3) 方法通量高, 费用低, 样品用量少。

3. 代谢组学发展历史及现状^[1, 2]

代谢组学研究可追溯到始于上世纪 70 年代的代谢谱分析, 这类代谢谱分析通常采用气相色谱技术 (GC, Gas Chromatography) 对患者体液中代谢物进行定性、定量分析, 从而对疾病进行筛选和诊断; 这种在临床上利用代谢谱分析诊断有关疾病的方法一直延用至今; 1983 年, 荷兰科学家 Van Der Greef 在国际上首先对尿液中代谢指纹进行质谱研究, 随后陆续有不少科学家开始应用高效液相色谱-质谱联用技术 (High Performance Liquid Chromatography / Mass Spectrometry, HPLC/MS) 和核磁共振 (Nuclear Magnetic Resonance, NMR) 技术进行代谢谱分析; 90 年代后, 代谢组学研究主要集中在药物的体内代谢等方向; 1997 年, Oliver 提出了通过定量分析尽可能多的代谢产物评估酵母基因的遗传功能及其冗余度的必要性, 首次将代谢产物和生物基因的功能联系起来; 1999 年, Nicholson 等提出 metabonomics 的概念, 2000 年, 德国马普所的 Fiehn 等提出了 metabolomics 的概念。通常, 植物和微生物领域应用色谱质谱技术进行的细胞代谢组学研究多采用 metabolomics, 而在药物研发和疾病研究等领域, 采用 NMR 技术对动物体液或组织样品进行研究的, 则用 metabonomics 较多。随着研究的深入, 现在这两个名词已基本等同使用。

目前, 代谢组学已广泛应用于药物研发、营养学、毒理学、临床化学等多个领域, 其中包括由于基因修饰引起的机体代谢效应的研究、先天性代谢系统缺陷研究、代谢系统相关的疾病(诸如糖尿病等)诊断与发病机制研究等, 均取得了一系列有意义的成果。

二、代谢组学最新研究进展

随着研究的不断深入, 越来越多的疾病发病机制, 以及机体代谢变化被揭示出来, 各种日新月异的方法技术手段也层出不穷, 为从代谢角度进行疾病的诊断和治疗打下了坚实的基础。 以下就代谢领域内几项最新的研究热点进行简单介绍:

1. 空间代谢组学

代谢物的合成和累积往往具有精准的空间分布, 且生理功能常与其在组织甚至单细胞中的空间分布紧密相关。因此, 实现高空间分辨地精准定位组织中代谢物的分布对阐明代谢物的合成、积累和调控机理至关重要。质谱成像 (MSI) 利用质谱测定分子质量并可视化特定化学成分 (如生物标志物、代谢物、肽/蛋白质) 的空间分布, 是一种很有前途的分析技术。

空间代谢组学是代谢组学和质谱成像技术 (MSI) 结合, 将组学信息扩展到二维乃至三维的水平, 研究小分子在组织切片中的空间分布, 极大拓展了对生物体中代谢物空间组成及分布的认知。比如以往的代谢组学寻找差异成分, 需要依靠峰面积 Ratio 比值。在代谢物含量差异不大的情况下, 难以找到生物标志物。而空间代谢组学, 可以在空间图像上直观的找出代谢物空间上的差异。以下将对常见的空间代谢组学技术, 包括基质辅助激光解析电离质谱成像技术 (MALDI-MSI)、解吸电喷雾电离质谱成像技术 (DESI-MSI) 以及空气动力辅助离子源质谱成像技术 (AFAI-MSI) 进行原理、特点以及应用方面的简要阐述。

1.1 基质辅助激光解析电离质谱成像技术

基质辅助激光电离 (MALDI, Matrix-Assisted Laser Desorption/Ionization) 是在真空中用激光照射样品与基质形成的共结晶薄膜, 基质从激光中吸收能量传递给生物分子, 而电离过程中将质子转移到生物分子或从生物分子得到质子, 而使生物分子电离的过程。它是一种软电离技术, 可以使大分子带上电荷而不过度碎裂, 适用于混合物及生物大分子的测定。

基质辅助激光解吸电离 (MALDI-MSI) 作为一种强大的分析工具, 具有高的空间分辨率和分子特异性, 有助于对内源性和外源性化合物进行无标记跟踪。与传统成像技术相比, MALDI-MSI 具有无需标记、灵敏度高、

分子特异性强、能一次从组织标本中同时原位定位多种生物分子等特点, 提供化合物的高通量扫描, 可实现物质的原位解析。

MALDI-MSI 已成为生物医学研究领域最具潜力的技术之一, 并在生物标记物发现, 药物开发, 疾病诊断和预后方面发现了其应用前景。比如, 作为一种高效且发展较为成熟的应用, MALDI-MSI 技术已经被广泛地用于植物代谢组学研究中^[1]。该技术无需对目标对象事先有确切的了解, 便可确定植物组织中存在的诸如碳水化合物、蛋白质、脂质等化合物的二维坐标, 以可视化代谢产物的空间分布, 这是这种方法的一个巨大优点。此外, 代谢物的鉴定不需要参考样品, 因为每个代谢物都可以通过其 m/z 值 (质核比) 进行鉴定。然而在研究植物组织时, 代谢物的识别和量化只是一部分, 研究它们如何分布甚至更为重要。这些结果可以用于鉴定含有最多目标次生代谢物的植物器官, 以便后续提取; 也可以用来确定代谢物提取的方法, 例如, 如果目标代谢物在液泡内, 则应采取细胞均质化技术对其进行提取; 对代谢产物分布的了解也可用于了解植物的一级和二级代谢机制, 进一步用于研究新的施肥方法或其他增产的方法。

1.2 解吸电喷雾电离质谱成像

解吸电喷雾电离质谱成像 (DESI-IMS, Desorption Electrospray Ionization-Mass Spectrometry Imaging) 利用电喷雾电离的原始原理, 在环境条件下, 将溶剂滴光栅化并直接在样品表面上解吸, 可以进行快速且破坏性小的化学筛选, 并能够以 μm 尺度的分辨率准确观察组织中的分子, 并且数据集可以提供每个检测到的分析物的图片。DESI-IMS 分析不需要复杂的样品制备。由于其简便的工作流程和多功能性, DESI-IMS 已成功应用于许多不同的研究领域, 例如临床分析, 癌症研究, 环境科学, 微生物学, 化学生态学和药物发现。

DESI-IMS 的主要应用领域是医学研究, 它用快速、准确的可视化方式运用在生物组织中的生物标志物发现。特别是, 已证明 DESI-IMS 能够直接从生物组织中获得大量已知或未知的内源性代谢物和外源性药物等分子的结构、含量和空间分布信息。通过优化的空气动力辅助解吸电喷雾质谱成像 (AFADESI-MSI) 技术 (是基于 DESI-IMS 技术上的一个升级, 能解决 DESI-IMS 鉴定出的代谢物少的瓶颈), 可绘制生物样本中超过 1000+ 种代谢物的空间分布图谱, 将功能代谢物的时空变化与组织结构和生物功能联系起来, 有助于分子组织学和分子病理学研究。在医学和临床研究中, AFADESI-IMS 被广泛的应用在疾病分子机制、生殖发育、肿瘤代谢与肿瘤免疫、肿瘤分子病理诊断、生物标志物筛选、药物药理和毒理等进行研究。

再帕尔·阿不力孜、贺玖明课题组^[4]对自主研发的气流辅助解吸电喷雾电离质谱成像 (AFADESI-MSI) DESI-MSI 的升级技术进行了优化, 通过非靶向分析, 在大鼠脑、肾脏和人食道癌组织中观察到数千种代谢物, 为分子组织学研究提供了有力的分析工具。研究通过优化后的 AFADESI-MSI 检测到了 1500 多种代谢物, 包括胆碱, 多胺, 氨基酸, 肉碱, 核苷, 有机酸, 碳水化合物, 脂质和其他小分子代谢物。同时该课题组通过 AFADESI-MSI 分析食管癌组织的冷冻切片, 结果表明精胺和亚精胺在癌组织中被上调, 对应肿瘤强大的增殖能力。谷氨酰胺作为癌细胞中主要的三羧酸 (TCA) 循环碳源也显示出在肿瘤组织中的大量消耗。该方法具有覆盖范围广, 灵敏度高, 动态范围宽和特异性强等优点, 在空间代谢组学研究和分子组织学分析中显示出广阔的应用前景。

1.3 空气动力辅助离子源质谱成像 (AFAI-MSI)

原位电离中的一个关键过程是带电样品液滴或离子通过大气压界面的有效收集和传输, 最终进入质谱仪。设计空气动力辅助电离系统 (AFAI) 的最初目的是通过高空气流量, 延长离子或带电液滴的传输距离以提高收集和传输的效率。

目前, 空气动力辅助离子源质谱成像 (AFAI-MSI, Air Flow Assisted Ionization) 是一种新型的质谱成像方法, 是表征生物组织脂沉积症的有力工具。在无需样品预处理或标记的条件下, 即可实现组织成像。AFAI-MSI 可以生成多色图来说明组织的空间分布和感兴趣分子的相对强度。这种技术产生多分子空间分布的可靠的图像, 在一个给定部分内的多分子的空间分布 AFAI-MSI 最重要的优点是保存了分子特性, 可以用来观察组织切片中脂质的不均匀分布。

此外, AFAI-MSI 是一种快速且近乎实时的分析方法。对组织切片进行一次 AFAI-MSI 分析只需数十分钟, 它已被成功用于实现高极性和非极性分子, 包括染料药物样品、炸药、滥用药物、蛋白质和挥发性化合物等的电离分析。而利用 AFAI-MSI 技术也实现了手指和全身分子成像的残留检测, 这些结果也表明 AFAI-MSI 可能是一种潜在的手术中切除工具。

2015 年, 贺玖明课题组利用 AFAI-MSI 技术对乳腺肿瘤细胞进行空间代谢组学分析, 建立了一种快速高效的方法以获取乳腺组织的分子信息, 可以在原位环境中通过分析脂质来鉴别乳腺癌, 并利用该方法鉴别了乳腺浸润性导管癌 (breast invasive ductal carcinoma, IDC) 和原为导管癌 (breast ductal carcinoma in situ, DCIS) 的各种亚型和组织学分型, 结果表明 IDC 的磷脂和脂肪酸含量均高于 DCIS。亚型和组织学分级验证集的标本分类与组织病理学诊断的一致性较高, 分别达到了 100% 和 78.6%。这项研究利用 AFAI-MSI 对乳腺癌的快速鉴别,

表明这种方法可以在外科手术中提供近乎实时的病人信息, 以对手术进度更好的把控。

2. 单细胞代谢组学

单细胞代谢组学 (Single cell metabolomics), 即在单个细胞水平上探索细胞代谢。采用单细胞代谢组学方法可以了解细胞间的细微差异 (即细胞异质性), 使研究结果与单个细胞或小细胞亚群相对应。单个细胞能够提供的用于分析的代谢物浓度低、体积小, 一些极为稀少的代谢物需要更加灵敏的检测方法。此外, 细胞内代谢物的浓度也可能在极短时间内因代谢周转 (metabolic turnover) 或环境变化等因素剧烈变化, 因此目前对于单细胞代谢组学的研究都采用灵敏度较高的质谱分析技术。

2.1 单细胞代谢组学应用

单细胞代谢组学可以应用于检测和分析癌细胞, 在正常代谢的细胞中发现具有异常高代谢速率的癌细胞, 检测出癌转移的循环肿瘤细胞。在目前的癌症治疗框架中, 利用单细胞代谢组学技术可以在癌组织中发现产生药物抵抗的癌细胞, 或者解释为何一些癌细胞在受到环境或药物刺激后会死亡而仍然还有一些细胞能够通过改变自身代谢途径存活下来。其他潜在应用还包括通过单细胞代谢组学技术获得建立细胞代谢数学模型所需要的输入和输出数据, 对衰老和干细胞命运进行更多的了解。

在单细胞代谢组学真正应用于系统生物学和医学诊断之前, 还有许多问题需要解决, 其中包括: 1) 扩大对代谢产物的覆盖范围; 2) 对单细胞代谢产物进行更快速的鉴定同时实现高通量检测; 3) 建立发现新的未知代谢产物的操作流程; 4) 对活细胞内的代谢产物进行检测。

2.2 单细胞代谢组学技术发展

在单细胞代谢组学研究中, 单细胞样品制备、细胞内代谢产物鉴定和数据分析都需要复杂的技术和模型来进行。

2.2.1 单细胞代谢组学样品制备^[5]

由于细胞代谢会对环境变化产生响应, 因此在单细胞样品制备过程中面临的一个主要问题就是如何在样品制备过程中尽量避免或减少对细胞代谢的影响, 其中一种方法就是在制备过程中将细胞尽可能地维持在天然环境中, 许多微流体芯片可以实现分离细胞后将其培养在合适的环境中, 向培养液中注射定量的化学物质, 有选择性地释放出细胞进行分析。另一种方法称为细胞淬灭, 细胞淬灭是指快速使细胞内的酶失活, 阻止代谢物的变化。因为单细胞的代谢物可以在短时间内发生巨大变化, 故对单细胞的代谢物检测来说, 细胞活性的淬灭是一个至关重要的问题。它可保证代谢物的较少变化, 并可避免产生错误的结果

2.2.2 提高单细胞代谢组学检测通量

研究人员通常使用质谱或核磁共振来开展代谢组学研究,但是由于核磁共振技术不太敏感,所以质谱法成为主要方法,现在已经开发出很多可以提高质谱检测能力和通量、简化提取单细胞代谢物的方法。

Zenobi 等人使用经特殊处理的硅载玻片,在添加稀释后的细胞后,递送到质谱仪里进行单细胞代谢组学分析。载玻片上涂层的排斥性确保每个储存空间只有一些液体和一到两个细胞,然后再借助仪器检测每个孔里细胞的代谢产物。这种芯片可以一次性检测上千个细胞,实现了较高通量的单细胞代谢组学研究。

2.2.3 单细胞代谢组学细胞破碎^[6]

由于动物细胞、植物细胞和微生物细胞在细胞大小上存在显著差异,而随着细胞大小的不同,细胞内代谢物的绝对量也存在变化。Akos Vertes 指出,对于比较大的细胞可以使用尖锐的光纤将红外光传递到细胞内,光的激发导致细胞炸裂并喷射出细胞内的生化分子,随后借助后续技术进行质谱分析,而通过计算机可以操控实现该过程的自动化;对于小细胞,可以将细胞沉积在硅材料制成的纳米涂层上,通过对涂层进行成像揭示离子束向哪个位置进行单细胞靶向,通过后续技术进行代谢物分子的质谱分析。

2.2.4 扩大代谢产物覆盖范围

目前常利用生物信息学方法解读单细胞代谢组学实验结果,主要挑战在于开发可以分析多个单细胞并检测每个细胞中的代谢产物、同时还可以使结果具有统计意义的仪器。目前的方法通常只能研究几十或上百个不同分子。但是细胞中存在许多种代谢物,即使是目前最敏感的分析技术,也只能检测细胞中最普遍也最容易检测的分子,不太常见的分子就难以被检测。除此之外,如何对单个活细胞进行代谢组学分析也是科学家们努力的方向。

3. 功能代谢组学^[7]

功能代谢组学是在传统的以发现差异代谢物为主要目的的非靶标、靶标代谢组学基础上,综合结合分子细胞生物学实验、同位素标记、宏基因组、转录组和蛋白组等多种技术手段,探究差异代谢物的生物功能以及相关生理病理意义的一种新方法,是传统代谢组学概念的深化和延伸,克服了传统发现代谢组学主要依靠文献进行生物学解释的缺点。以下主要对功能代谢组学技术平台与应用前景做出简要介绍。

3.1 代谢组学技术平台

传统的发现代谢组学研究给出了潜在的生物标志物、紊乱的代谢途径的结果,但这些研究结果的进一步分析解释主要是基于相关文献的解读。为了功能探索和

代谢组学数据的进一步验证,需要包括体内和体外实验在内的细胞生物学研究平台。

3.1.1 基因操作平台——构建稳定表达细胞系

为了研究基因改变对代谢的影响,在功能代谢组学研究中需要稳定表达的细胞系。往往涉及到基因的敲除,敲入等基因编辑手段。近年来,CRISPR/Cas9 技术成为一种强大的基因编辑工具,可以长期稳定保持性状,成为功能代谢组学研究中热门手段^[8]。基因编辑可以迅速改善细胞的一系列性状,在进行基因编辑后使用下游组学(如代谢组学)对其代谢情况进行研究,可以探究验证所编辑基因的功能,服务于上游组学。

3.1.2 分子生物学平台——研究代谢产物及相关蛋白的功能

细胞中的代谢物具有许多功能,如修饰蛋白质和基因、作为信号分子、提供细胞能量等。某些代谢物可促进疾病进展,例如,2-羟基戊二酸(2-HG)被认为是胶质瘤中的一种肿瘤代谢物^[9];富马酸被认为是肾癌的一种肿瘤代谢产物^[10]。除这些代谢物外,还有许多其他的代谢物也具有显著的促进或抑制疾病的作用,但更多的分子功能有待阐明。

在功能研究方面,需要一个全面的分子生物学平台来帮助我们了解代谢产物在疾病病理生理学中的作用,例如,测试代谢产物的改变是否影响癌细胞的增殖,细胞是否应进行增殖试验。最近,海马分析仪成为了一个强大的工具。它在许多研究中被用来分析细胞中的代谢流量^[11, 12]。它可以提供细胞外酸化率(ECAR)和氧消耗率(OCR),以提供糖酵解和线粒体呼吸效率的信息,被视为代谢组学分析的有力验证。

3.1.3 系统生物学平台——整合多组学数据

我生物体是一个多层次调控系统,单一的组学数据只能提供不完整的、有偏倚的部分信息。因此在探索代谢产物的研究中,往往需要将基因组学、转录组学或蛋白质组学与代谢组学数据相结合进行分析。

系统生物学是后基因组时代的新兴学科,旨在对生物体进行系统的研究。随着组学技术的成熟和大数据处理能力的提高,系统生物学研究日趋明朗。然而我们对于基因的功能了解还甚少,尤其是涉及到基因-蛋白质-代谢产物调控网络。

整合多组数据可以重建多层调控网络,提供基因对转录本、转录本对蛋白质、蛋白质对代谢物等更全面的调控信息。更重要的是,不同层次的组学数据可以相互验证。代谢组学与转录组学的结合可用于研究疾病机制^[13]。而由生物信息学专家们编写的多组学数据集成软件,如 Integrator、MultiDataSet、MPLEx 等,可以更为高效地处理分析多组学数据。

3.2 功能代谢组学应用前景

目前功能代谢组学在癌症、代谢性疾病和生物表型研究中已被广泛应用, 在应用中表现出具有巨大优势和广阔发展前景。

3.2.1 癌症的检测

1924 年, Warburg 效应的发现引起了癌症代谢研究科学家的广泛关注, 癌细胞大量消耗葡萄糖和谷氨酰胺, 且其代谢速率异常快。偏好利用有氧糖酵解而不是氧化磷酸化产生 ATP 被认为是癌症特有的特征。目前, 有说法认为癌症是一种代谢性疾病, 因为相关研究表明在癌症的发生和发展过程中会发生代谢重编程^[14]。而功能代谢组学可以高效地发现与癌症相关的代谢改变, 提供对肿瘤样本中数千种代谢物的全面分析。如今代谢组学已广泛应用于癌症研究。

由于肿瘤的异质性, 肿瘤的代谢非常复杂。胰腺导管癌 (Pancreatic ductal carcinoma, PDAC) 是一种难以治愈的癌症, 通常诊断为晚期疾病。早期发现并干预 PDAC 可有效提高患者生存期。通常对患者胰腺组织血浆代谢分析, 发现 BCAA (直链氨基酸, 蛋白分解增加的标志) 的含量上调, 提示患者可能处于 PDAC 早期阶段。使用功能代谢组学进行肿瘤标志物检测, 具有巨大的潜在价值^[15]。

3.2.2 代谢性疾病的检测与治疗

代谢性疾病是指由代谢紊乱引起的一种疾病, 包括糖尿病、高脂血症、高血压、代谢综合征、非酒精性脂肪肝等。代谢性疾病的发生和发展与基因改变、不合理的饮食习惯和环境影响有关。虽然代谢性疾病在世界范围内造成了许多人死亡, 但其发病机制尚不明确。

代谢性疾病患者常伴有代谢变化。为了研究这些代谢变化, 功能代谢组学是一种很好的研究方法。为了预测 II 型糖尿病的发生, Wang 等人采用基于 LC-MS 的高通量代谢谱研究糖尿病相关的代谢风险, 对 2422 名血糖正常的人进行了 12 年的随访, 其中 201 人发展成糖尿病。结果表明 BCAA、酪氨酸和苯丙氨酸的代谢产物可能与糖尿病风险相关, 可用于预测糖尿病^[16]。此外, Lee 等人在白种人和西班牙裔人群中进一步证明了 BCAA 与糖尿病的强烈相关性^[17]。

3.2.3 生物表型的研究

在所有组学中, 代谢组学是最接近表型的组学, 故而利用代谢组学研究生物表型是一种很好的方法。研究代谢-表型相互作用有助于我们更好地了解生物体内的调控网络。

蝗虫分为独居型和群居型。群居蝗虫对农业危害更大。然而, 这两种表型的代谢调控尚不清楚。有研究进行了基于 LC-MS 的代谢组学研究, 发现肉碱和酰基肉碱的变化与蝗虫的相变有关。此外, 该团队也发现通过沉默这两种代谢产物的关键酶, 可以实现蝗虫的相变^[18]。

基于每个人基因型都是独特的, 精准医疗旨在实现个性化的治疗。对于精准医学研究, 了解个体的特异性是非常重要的。代谢组学也可以作为精准医疗的工具。例如, 代谢组学相关技术已被用于研究长期服用神经抑制剂 18 个月的精神分裂症患者, 结果表明许多代谢物的变化与表型有关^[19]。此外, 代谢组学还可用于预测健康个体的疾病。Guo 等人结合代谢组学与全外显子组测序, 对 80 名健康志愿者进行了代谢组学研究。结果发现代谢物的改变具有个体特异性, 且与基因改变有关, 这一发现表明功能代谢组学可用于疾病风险评估^[20]。

4. 代谢流组学^[21]

常规代谢组学检测的是静态的代谢物浓度, 但是静态的含量有时候不能完全说明问题, 需要结合动态分析予以更为全面的判断。例如街道上的交通量会随着车辆密度的增加而增加, 导致交通拥堵发生。但道路上车辆的高密度并不能解释交通量大的原因, 真正的原因在于道路出入口车流流量存在差异。同样, 在新陈代谢中, 代谢物的积累可能是由于产物生成的增加, 也可能是产物消耗的减少导致的。因此需要代谢流技术介入研究, 进行动态的代谢研究。

代谢流是利用稳定同位素示踪技术析下游代谢产物的同位素标记模式, 从而推算出该化合物在代谢通路中的流向和分布; 通过对生物体进行代谢流分析, 可得到生物体特定代谢通路的活跃程度。当一个代谢物产生积累时, 可能是由于其生产的增加或者是消耗的减少。基于稳定同位素示踪法的 MFA 则可以帮助我们测量代谢流量: 带有稳定同位素标记的代谢物经过生化反应, 则会导致下游代谢产物的标记, 产生在特定位置被同位素标记的质核比分别为 M+1, M+2,M+n 等的代谢物。通过分析下游代谢物的标记模式及被标记代谢物的量, 由此可以计算得出感兴趣的代谢通路的流量速度和方向信息。

4.1 基于质谱的代谢流分析流程和原理

完整的代谢流分析流程一般包括: 稳定同位素标记样本的制备、色谱-质谱检测、标记和非标记代谢物定量、天然同位素丰度校正和计算标记特征等几个步骤。主要创新之处在于引入同位素标记对代谢过程进行动态研究。下对关键步骤做简要概述。

4.1.1 稳定同位素标记

将目标代谢产物分子标记上稳定性同位素可以观察其在代谢过程中随时间动态变化的过程。根据实验目的, 选择合适的稳定同位素标记底物, 通过更换标记培养基或注射到动物体内对下游代谢产物进行标记, 之后体系达到代谢稳态。如[U-13C]glucose 可流入中心碳代谢途径 (EMP/TCA/PPP) 及其分支途径; [U-13C]glutamine 可流入 TCA 循环、谷氨酰胺和谷氨酸代

另外样本收集的时候要通过快速淬灭代谢酶来获得更准确的实验结果,一般可以通过液氮速冻或者快速加入低温预冷的挥发性有机溶剂来实现。

4.1.2 天然同位素矫正

天然稳定同位素可能会使目标代谢物质量数更高,这在后续分析中会带来误差。比如,测量得到的 M+1,有可能是 1 个碳原子被 ¹³C 标记,也有可能是 1 个氢原子被 ²H 标记,或者 1 个氧原子被 ¹⁷O 标记。因此在对分析代谢组学数据结果时需要元素的天然同位素丰度进行数据矫正,以获得真实的质量分布向量 (Corrected MDV)。

质量分布向量 (Mass Distribution Vector, MDV) 也叫做质量同位素分布向量 (Mass Isotopomer Distribution Vector, MID vector), 描述的是标记模式下代谢物的每个同位素分布的比例。不同同位素标记形式 (M+1, M+2...) 反映了通过不同代谢途径得到的产物量,通过计算 MDV 可以帮助比较不同代谢通路的活性。

4.2 代谢流组学应用

代谢流技术可以帮助更好地理解细胞内代谢网络的代谢物水平变化、流量分布和周转速率,发掘主要代谢异常通路及其生物学功能,并揭示其上下游相互调控机制;可为疾病发生机制的理解、药物靶点的发现等提供强有力的科学依据。

4.2.1 代谢通路示踪

代谢是高度复杂且受严密调控的动态变化网络。除了基于特定酶、转运体的调控外,通路之间可以通过同一中间产物而产生关联。如果能找到肿瘤细胞中相较于正常细胞而特定依赖的代谢通路,那么我们就可以精确地靶向肿瘤细胞进行治疗和干预。

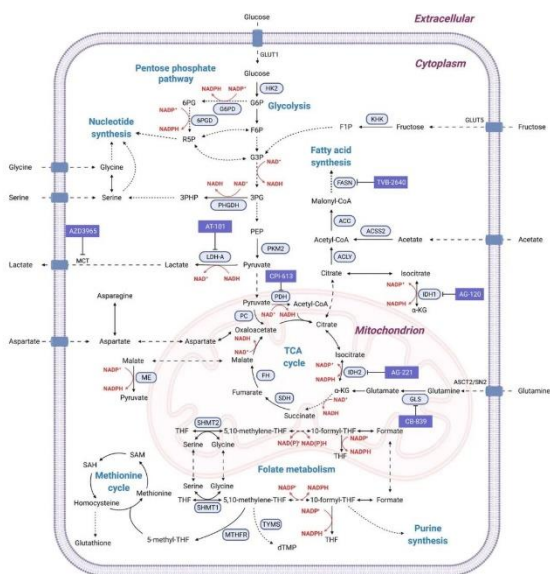


图 1 促进肿瘤细胞生长的代谢通路及潜在治疗靶点

图 1 展示了细胞中复杂的代谢通路,包括葡萄糖的代谢(糖酵解、磷酸戊糖途径)、脂肪酸代谢、核苷酸的合成、叶酸代谢等,其中特别标记了值得调控的关键酶和转运体,以及针对这些作为靶标已进入临床试验或者已经被 FDA 批准的小分子药物。譬如,在胶质瘤中曾报道过突变的异柠檬酸脱氢酶 (IDH) 可以介导肿瘤代谢物 2-羟戊二酸 (2HG) 的产生,展示了 IDH 作为抗肿瘤靶标的潜力,从而引发 IDH 抑制剂的开发、获批与应用。

4.2.2 肿瘤药物研发

代谢组学与代谢流分析也可以在肿瘤药物研发中发挥重要作用,并可贯穿于每一步中:从发现靶点到理解药物作用机理,从耐药机制研究到指导精准治疗。

经过代谢组学分析后,差异代谢物和代谢通路可引导发现潜在的生物标记物和可靶向的代谢依赖性和弱点。潜在的生物标记物可帮助肿瘤的早期诊断、预后和药物有效性预测。通过结合代谢流分析,代谢靶标可以帮助新药研发,或者是帮助科研人员更好地理解现有药物的作用机制,以及如何产生耐药,从而改善现有疗法。药理代谢组学可以用于指导精准治疗;饮食干预法则可以作为药物治疗的辅助手段。

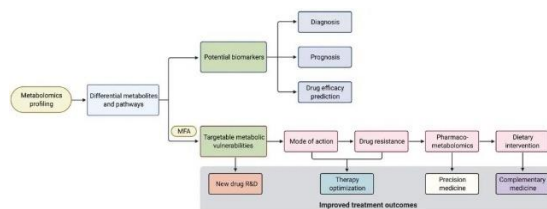


图 2 代谢组学和代谢流分析技术在肿瘤药物研发和药理学中的应用

尽管代谢组学和代谢流分析极大拓展了我们对于肿瘤生物学的理解,但是领域中依旧存在诸多技术挑战和瓶颈,比如灵敏度不足、精准度不够、难以进行代谢流分析,以及至今无法实现真正意义上的单细胞代谢组学(特别是由于灵敏度的技术瓶颈)等等。相关的技术进步和新型方法开发都将进一步促进代谢组学和代谢流分析技术在不同生物医学背景下的应用。下一阶段的研究需要更好地整合、利用所获取的代谢重塑表型和机制信息,将其转化成更好的抗肿瘤疗法。药物研发方面需要更多地关注肿瘤微环境,尤其是肿瘤细胞与免疫细胞之间的代谢相互作用。多组学整合的应用,包括基因组学、蛋白组学、代谢组学等,将有助于加深我们对于肿瘤生物学的理解和利用,进一步加速抗肿瘤药物的研发。

5. 深度学习在代谢组学中的应用^[22]

随着代谢组学的飞速发展,研究人员能够获得大量高灵敏度和高分辨率的代谢组学数据。复杂生物信息的获取和解释对传统的数据分析策略(即功能表征、注释

和集成)提出了更高的要求,从而增加了对开发新的计算工具的需求。

深度学习(Deep learning, DL)一词最初在 1986 年被引入机器学习(machine learning, ML),后来在 2000 年时被用于人工神经网络(Artificial neural networks, ANN),是机器学习和人工智能(artificial intelligence, AI)的一个新兴研究领域^[23]。深度学习(DL)由多个层组成,以学习具有多个抽象层次的数据特征。通过多层处理,逐渐将初始的“低层”特征表示转化为“高层”特征表示后,用“简单模型”即可完成复杂的分类等学习任务。由此可将 DL 理解为进行“特征学习”(feature learning)或“表示学习”(representation learning)。DL 可以从非结构化的、不同的数据集中学习并预测关系。DL 采用 DNNs (“深度”神经网络)获取输入数据,并通过依次结合前一层的输出,将其转换为抽象的特征表示,使这些神经网络成为“深度”神经网络的特征(图 4)。DL 模型的复杂结构允许对大型和/或高维数据集所显示的复杂模式进行建模。各种 DL 模型已经得到了广泛的应用。与传统的 ML 模型相比,DL 模型提高了数据分析的能力和可解释性,在改善疾病预测、诊断和药物发现等方面具有巨大的潜力。

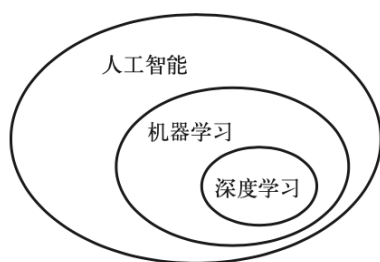


图 3 人工智能、机器学习与深度学习

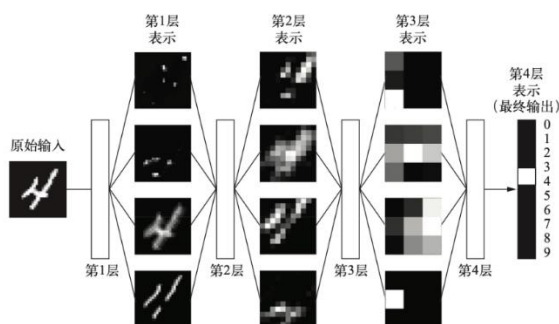


图 4 数字图像分类模型学到的深度表示

5.1 深度学习(DL)在代谢组学数据采集和处理方面的应用

近年来, DL 已逐渐被应用于质谱的数据采集和处理,如噪音过滤、峰检测、整合和对齐、重叠峰的去卷积、子结构预测和代谢物识别。

5.1.1 峰检测与分类——ANN/Peakonly/Wipp

峰检测在实际应用中被视为一种模式识别问题,可采用了一种人工神经网络(ANN)方法来解决^[24, 25]。与

经典的基于回归的方法相比, ANN 方法估计了一个特定峰的加权检测概率,并在最终的数据集中进行取舍。但是,这种方法成本很高,并且受到计算峰面积或反卷积重叠峰能力的限制。

为了识别 LC-MS 代谢组学数据中缺失的峰,有人开发了一种卷积神经网络(CNN)方法——Peakonly^[26]。Peakonly 可以同时排除低强度(噪音)峰,保留真正的正峰。但这种方法尚未与其它多变量统计和 ML 方法,如 PLS-DA (偏最小二乘法判别分析)、ANN、加权回归模型进行比较。Kantz 等人,采用 DNN (“深度”神经网络)和 MLR(多元逻辑回归)联合方法对非靶向代谢组学 LC-MS 数据的质谱特征进行分类。该方法可去除 90% 的假阳性峰(噪音),并保留真阳性峰。但这种方法可能需要对实验室中每一种新建的 LC-MS 方法进行训练和验证。因为该方法本身是将峰的形状转换为图形对象,每个真正的峰会根据特定的分析条件而变化。使用 MLR 算法模型可克服其中一些问题,因为它允许直接从峰形状获得的参数,但这些参数仍然需要根据具体的分析条件进行调整。

WiPP 是一种基于 ML 的多算法工作流程,用于 GC-MS 中峰检测^[27]。WiPP 可以通过结合 7 个峰分类器的分类方案来评估检测到的峰的质量,为代谢物识别和下游分析提供检测特征。

5.1.2 GC-MS 峰对齐——ChromAlignNet

除了峰的检测和分类外, DL 还被应用于 GC-MS 的峰对齐。ChromAlignNet 是一种旨在改进使用高分辨质谱[TOF-MS]获得的数据峰对齐的方法^[28]。该技术使用循环神经网络(RNN)来对单个峰(特定的 m/z 和保留时间)水平上的数据进行训练和学习。通常,该算法检测出真阳性结果可信度(84-96%)很高。但在应用于呼吸样本时,它假阳性率高达 30%,表明,需要更多的工作来减少从 GC-MS 数据中检测到的假阳性峰。

5.1.3 串联质谱识别代谢物——DeepMass

使用串联质谱(MS/MS)来识别“未知”代谢物(没有参考谱和/或结构信息的代谢物)存在一些问题,因为其参考质谱图数量有限,并且化合物的质谱图可能随仪器类型和所施加的碰撞能量而变化。DeepMASS 是一种基于 DNN 的开源框架,被设计用来识别“未知”代谢物^[29]。它主要有以下三个关键步骤:(1)使用 MS/MS 谱数据集训练一个 DL 模型;(2)对“未知”化合物与 MS/MS 谱数据库的结构相似性进行评分;(3)对候选代谢物进行排序。

5.1.4 预测代谢物电子-电离质谱结果——NEIMS/Canopus/MetDNA/NormAE

NEIMS 是一种被设计用来预测给定代谢物的电子-电离质谱的神经网络(NN)算法,可为数千个可能的候选者生成质谱^[30]。通过预测 NIST 质谱库中小分子的质谱来测试 NEIMS 的性能。结果表明, NEIMS 方法相对较

快(每个分子 5ms, 准确率为 91.8%), 但这种方法并没有考虑到同位素峰的强度。

Duhrkop 等人^[31], 开发了 Canopus, 一个基于 DNN 的算法, 从质谱碎片中预测化合物类别。在 LC-MS/MS 运行中, Canopus 可以为每个代谢物(包括“未知物”)分配类别信息。在该研究中, 使用模拟结构数据库中的化合物结构训练一个 DL 模型, 并使用分子指纹预测它们的类别。此外, 还对“未知”类的代谢物进行了注释。通过该算法以及一个内核 SVM 方法, 研究了小鼠消化系统的化学多样性和微生物定植情况, 并推断在化合物类水平上发生的生物反应。

MetDNA 是另一种代谢物注释和异常调节分析算法, 通过确定一个代谢网络(具有相似的质谱)的“反应对”中未知化合物的结构相似性来假定地注释了“未知的”代谢物^[32]。

NormAE 是一种旨在消除在大规模代谢组学研究中非靶向 LC-MS 数据中经常出现的非线性批处理效应新 DL 技术^[33]。它是基于非线性自动编码器(AEs)和对抗性学习技术的。与其他传统的归一化方法相比, NormAE 的校准结果最好。

5.2 深度学习(DL)在代谢表型的分层中的应用——寻找代谢特征

代谢组学实验往往能检测到数百乃至数千种代谢物, 获得比现有样本更多的特征。尽管代谢组学数据具有非线性性质, 但用于“代谢分型”的非线性 ML 方法尚未被广泛使用。这可能是因为线性方法(如 PLS-DA)比非线性 ML 方法更容易解释。此外, 由于代谢组学数据集的样本量较低, ANN 或 DNN 模型往往会过拟合。目前, 当经典的 PLS 方法无法有效发挥功能时, 越来越多的研究使用 DL 分析代谢谱并提取具有生物学意义的信息。

在分析代谢组学数据时, DL 可以捕获复杂性状的代谢特征。2020 年, Alakwaa 等人应用 DL 准确预测了乳腺癌样本中的雌激素受体状态。使用 DL 分析包含 162 种代谢物的数据集, 并与传统的 ML 方法(如 RF、SVM、PAMs、RPART 和 LDA 等)进行比较。结果表明, 与其它 ML 方法相比, DL 准确性最高(AUC~0.93)。此外, 这种 DL 方法还揭示了 8 条可能促进乳腺癌的新代谢途径^[34]。

DNN-MDA 是一种综合监督分类和回归技术, 用于样本分类和代谢物选择^[35]。与其它传统 ML 方法相比, DNN-MDA 方法分类准确率最高(97.8%), 但分类性能会随样本量呈线性变化。此外, 在分析高度偏倚数据集时, 预测能力会降低。Asakura 等人开发了一种集成 DNN(EDNN)算法, 以提高分类和回归性能。EDNN 集成了几种 DNN 分类器和统计方法, 如引导重采样、变量随机抽样, 并将不同分类器的结果进行组合。与 PCA、

RF、SVM 和经典 DNN 相比, EDNN 能更好的预测结果。

在另一项研究中, 将最近开发的两种 ML 算法包括 DL 和极端梯度增强(XGBoost)与 RF 在监督分类中进行了比较^[36]。通过该算法捕获了阿尔茨海默氏病(AD)的代谢复杂性, DL 与其他 ML 算法相比, 表现出相近的性能(AUC=0.85)。

DL 算法是非结构化数据集(例如成像数据)的一个巨大突破, 在基于成像的代谢组学研究中也显示出了突出的表现。Inglese 等人^[37]在一项基于 MS 的成像研究中使用 DL, 揭示了癌症的代谢异质性, 且开发了一个新的计算工作流程, 成功地识别了肿瘤组。此外, 他们还发现了在肿瘤组织中发生的新的化学和生物相互作用。

5.3 深度学习(DL)在多组学数据联合分析方面的应用

“多组学”联合分析也是目前的一个研究热点。从基因组到转录组, 再到蛋白质组和代谢组, 更多的“组学”联合分析为生命活动的研究提供了更丰富、详细的信息, 但是这种分析不可避免地会带来维数增加, 特征冗余, 以及不同的“组学”技术所伴随的信噪比。面对这种情况, 传统的 ML 方法不再充分实用。

在多组学分析中, DL 在单个组学数据集中提取它们的内部抽象特征, 消除了手工特征提取, 并提高了信噪比, 成功地将不同的组学技术结合在一起研究生物学问题。此外, 使用 AEs 和各种形式的神经网络, 能够成功地整合转录组、基因组甚至化学结构数据, 以预测抗癌化合物的敏感性。

有研究报道, DL 可通过整合来自妊娠中期的羊水蛋白质组学和代谢组学数据, 以及超声、临床和人口统计学数据, 预测无症状的短宫颈长度孕妇的围产期^[38]。此外, DL 模型在预测与分娩和/或新生儿重症监护病房相关的几种结果方面优于其他六种传统 ML 技术(如 SVM、GLM、PAM、RF 和 LDA)。Kim 等人将 4 种独立的“组学”数据(表型、蛋白质组学、代谢组学和基因组学)进行整合, 模拟全基因组影响大肠杆菌下游代谢和随后生长过程^[39]。

5.4 深度学习(DL)在代谢途径预测中的应用

目前通过 DL 还可预测许多与化合物相关的代谢途径。Li 等人, 开发了 mummichog 算法^[40], 无需预先识别, 即可直接从质谱特征表成功地预测代谢途径的功能。

AlAkwa 等人^[41]开发了 Lilikoi, 可以使用代谢组学数据进行代谢突进预测。与传统的通路工具(MetPA、IMPALA 和 MPEA, 它们使用代谢物名称进行通路映射、整合和富集分析)相比, Lilikoi 可以转换和合并样本中的代谢组学谱, 生成代谢途径。此外, 它还能指出一些与特定疾病表型显著相关的代谢途径。

Baranwal 等人^[42], 提出了一种混合 DL 架构和集成学习方法, 该方法结合了图卷积网络(GCN)和 RF 分类器来提取分子结构, 进行代谢路径预测。基于 GCN 的方法能够预测代谢途径, 准确率可达 95%, 此外, 这种方法还能估计代谢物对参与的多种代谢途径的相对贡献。

5.5 深度学习(DL)在人体代谢途中的应用

在过去的十年中, 随着高通量分子分析数据库的发展, 人类代谢组数据库、血清代谢组数据库、METLIN、REACTOME、KEGG、小分子通路数据库和人类周期数据库等数据框也在快速增长。

由于人类代谢研究产生的数据量不断增加, 因此需要开发新的计算工具来研究潜在的途径及其调控。DL 依靠大量的生物数据迅速发展, 已经开始在人类代谢组学的进展中发挥关键作用。

DL 的实用性可能不仅局限于以自动化的方式有效地分析代谢物和代谢通路, 而且可能对精准医疗有很大的价值。DL 已经在疾病诊断模型、药物发现和人类疾病的生物标志物的识别等方面显示出了巨大的潜力。例如, 一个基于慢性肾病(CKD)血清代谢组学数据 ML 模型能够正确识别 G1-G4CKD 患者 (93%)^[43]。同样, 基于深度 GCNs 的 DL 结构也被设计用于预测代谢途径、分子结构和特征, 准确率超过 90%。最近用于非靶向代谢物分析的 DL 的实现主要集中在在 MS 中使用高通量覆盖数据独立的获取上。这提高了基于 MS 的代谢物鉴定技术的准确性和稳定性, 表明 ML 在人类代谢物研究中具有很好的作用。

DL 已被证明在疾病表型、患者分类、新疗法的发现等方面具有重要的作用, 但其在代谢组学研究中的应用仍处于起步阶段。现阶段代谢组学研究中应用 DL 技术的样本量低、缺乏可解释性以及普遍缺乏足够的参考资料、训练和验证数据。此外, DL 模型还需要进一步被简化, 使其更易懂, 便于研究人员使用。

综上, DL 在解决代谢组学数据采集、处理、分析和解释等方面具有很大的潜力。结合 DL 和代谢组学, DL 可在医学诊断、生物标志物识别和药物发现等方面发挥重要作用。

三、参考文献

1. 许国旺, 路. 鑫, and 杨胜利, *代谢组学研究进展*. 中国医学科学院学报, 2007. **29**(6).
2. 马丽华, 杨宏静, and 徐晓艳, *代谢组学研究进展*. 现代医药卫生, 2017. **33**(17): p. 4.

3. Susniak, K., et al., *Recent developments of MALDI MSI application in plant tissues analysis*. Acta Biochim Pol, 2020. **67**(3): p. 277-281.
4. He, J., et al., *A Sensitive and Wide Coverage Ambient Mass Spectrometry Imaging Method for Functional Metabolites Based Molecular Histology*. Adv Sci (Weinh), 2018. **5**(11): p. 1800250.
5. Lipiec, E., et al., *Preparation of Well-Defined DNA Samples for Reproducible Nanospectroscopic Measurements*. Small, 2016. **12**(35): p. 4821-4829.
6. Lin, T.H., et al., *Low-Temperature Wet Conformal Nickel Silicide Deposition for Transistor Technology through an Organometallic Approach*. ACS Appl Mater Interfaces, 2017. **9**(5): p. 4948-4955.
7. Yan, M. and G. Xu, *Current and future perspectives of functional metabolomics in disease studies-A review*. Anal Chim Acta, 2018. **1037**: p. 41-54.
8. Fraser, P.D., et al., *Metabolomics should be deployed in the identification and characterization of gene-edited crops*. Plant J, 2020. **102**(5): p. 897-902.
9. Patay, Z., et al., *Cerebral neoplasms in L-2 hydroxyglutaric aciduria: 3 new cases and meta-analysis of literature data*. AJNR Am J Neuroradiol, 2012. **33**(5): p. 940-3.
10. Osu, N., et al., *Metabolic Alteration in Cancer Cells by Therapeutic Carbon Ions*. Anticancer Res, 2021. **41**(12): p. 6023-6029.
11. Kishton, R.J., et al., *AMPK Is Essential to Balance Glycolysis and Mitochondrial Metabolism to Control T-ALL Cell Stress and Survival*. Cell Metab, 2016. **23**(4): p. 649-62.
12. LeBleu, V.S., et al., *PGC-1 α mediates mitochondrial biogenesis and oxidative phosphorylation in cancer cells to promote metastasis*. Nat Cell Biol, 2014. **16**(10): p. 992-1003, 1-15.
13. Zhang, G., et al., *Integration of metabolomics and transcriptomics revealed a fatty acid network exerting growth inhibitory effects in human pancreatic cancer*. Clin Cancer Res, 2013. **19**(18): p. 4983-93.
14. Pavlova, N.N. and C.B. Thompson, *The Emerging Hallmarks of Cancer Metabolism*. Cell Metab, 2016. **23**(1): p. 27-47.
15. Mayers, J.R., et al., *Elevation of circulating branched-chain amino acids is an early event in*

- human pancreatic adenocarcinoma development*. Nat Med, 2014. **20**(10): p. 1193-1198.
16. Wang, T.J., et al., *Metabolite profiles and the risk of developing diabetes*. Nat Med, 2011. **17**(4): p. 448-53.
17. Lee, C.C., et al., *Branched-Chain Amino Acids and Insulin Metabolism: The Insulin Resistance Atherosclerosis Study (IRAS)*. Diabetes Care, 2016. **39**(4): p. 582-8.
18. Wu, R., et al., *Metabolomic analysis reveals that carnitines are key regulatory metabolites in phase transition of the locusts*. Proc Natl Acad Sci U S A, 2012. **109**(9): p. 3259-63.
19. Poldrack, R.A., et al., *Long-term neural and physiological phenotyping of a single human*. Nat Commun, 2015. **6**: p. 8885.
20. Guo, L., et al., *Plasma metabolomic profiles enhance precision medicine for volunteers of normal health*. Proc Natl Acad Sci U S A, 2015. **112**(35): p. E4901-10.
21. Liang, L., et al., *Metabolomics, metabolic flux analysis and cancer pharmacology*. Pharmacol Ther, 2021. **224**: p. 107827.
22. Sen, P., et al., *Deep learning meets metabolomics: a methodological perspective*. Brief Bioinform, 2021. **22**(2): p. 1531-1542.
23. Heaton, J., Ian Goodfellow, Yoshua Bengio, and Aaron Courville: *Deep learning*. Genetic Programming and Evolvable Machines, 2018. **19**(1): p. 305-307.
24. Mendez, K.M., D.I. Broadhurst, and S.N. Reinke, *Migrating from partial least squares discriminant analysis to artificial neural networks: a comparison of functionally equivalent visualisation and feature contribution tools using jupyter notebooks*. Metabolomics, 2020. **16**(2): p. 17.
25. Grapov, D., et al., *Rise of Deep Learning for Genomic, Proteomic, and Metabolomic Data Integration in Precision Medicine*. Omics, 2018. **22**(10): p. 630-636.
26. Melnikov, A.D., Y.P. Tsentalovich, and V.V. Yanshole, *Deep Learning for the Precise Peak Detection in High-Resolution LC-MS Data*. Anal Chem, 2020. **92**(1): p. 588-592.
27. Borgsmüller, N., et al., *WiPP: Workflow for Improved Peak Picking for Gas Chromatography-Mass Spectrometry (GC-MS) Data*. Metabolites, 2019. **9**(9).
28. Li, M. and X.R. Wang, *Peak alignment of gas chromatography-mass spectrometry data with deep learning*. J Chromatogr A, 2019. **1604**: p. 460476.
29. Ji, H., et al., *Deep MS/MS-Aided Structural-Similarity Scoring for Unknown Metabolite Identification*. Anal Chem, 2019. **91**(9): p. 5629-5637.
30. Wei, J.N., et al., *Rapid Prediction of Electron-Ionization Mass Spectrometry Using Neural Networks*. ACS Cent Sci, 2019. **5**(4): p. 700-708.
31. Dührkop, K., et al., *Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra*. Nat Biotechnol, 2021. **39**(4): p. 462-471.
32. Shen, X., et al., *Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics*. Nat Commun, 2019. **10**(1): p. 1516.
33. Rong, Z., et al., *NormAE: Deep Adversarial Learning Model to Remove Batch Effects in Liquid Chromatography Mass Spectrometry-Based Metabolomics Data*. Anal Chem, 2020. **92**(7): p. 5082-5090.
34. Alakwaa, F.M. and M.G. Savelieff, *Bioinformatics Analysis of Metabolomics Data Unveils Association of Metabolic Signatures with Methylation in Breast Cancer*. J Proteome Res, 2020. **19**(7): p. 2879-2889.
35. Asakura, T., Y. Date, and J. Kikuchi, *Application of ensemble deep neural network to metabolomics studies*. Anal Chim Acta, 2018. **1037**: p. 230-236.
36. Stamate, D., et al., *A metabolite-based machine learning approach to diagnose Alzheimer-type dementia in blood: Results from the European Medical Information Framework for Alzheimer disease biomarker discovery cohort*. Alzheimers Dement (N Y), 2019. **5**: p. 933-938.
37. Inglese, P., et al., *Deep learning and 3D-DESI imaging reveal the hidden metabolic heterogeneity of cancer*. Chem Sci, 2017. **8**(5): p. 3500-3511.
38. Bahado-Singh, R.O., et al., *Artificial intelligence and amniotic fluid multiomics: prediction of perinatal outcome in asymptomatic women with short cervix*. Ultrasound Obstet Gynecol, 2019. **54**(1): p. 110-118.

39. Kim, M., et al., *Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli*. Nat Commun, 2016. **7**: p. 13090.
40. Li, S., et al., *Predicting network activity from high throughput metabolomics*. PLoS Comput Biol, 2013. **9**(7): p. e1003123.
41. Alakwaa, F.M., K. Chaudhary, and L.X. Garmire, *Deep Learning Accurately Predicts Estrogen Receptor Status in Breast Cancer Metabolomics Data*. J Proteome Res, 2018. **17**(1): p. 337-347.
42. Baranwal, M., et al., *A deep learning architecture for metabolic pathway prediction*. Bioinformatics, 2020. **36**(8): p. 2547-2553.
43. Yang, Y., W.W. Tong, and Z. Liu. *A Machine Learning Model to Characterize Chronic Kidney Disease with Metabolomics Data*. in *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. 2019.