

第1章 绪论

1.1 引言

分子结构表征 (molecular structural characterization ,简称 MSC)是当代药物分子设计与药效评价中不可缺少的重要部分。MSC 是药物化学、有机化学等多个学科领域中一个古老而又年轻的课题,其手段随科学的发展与技术的进步而不断地更新、不断地完善。由于新分子的不断发现、不断创造,特别是组合化学技术的出现,使得在短时间内合成成千上万个分子成为可能。然而,分子结构,尤其是三维分子结构的实验测试技术进展缓慢,已越来越难以跟上新分子爆炸式增长的速度,数以万计的分子结构等待测试。近年基因组学、蛋白质组学与糖组学的相继兴起,大量核酸、蛋白质和糖等生物大分子的结构特别是高级结构(三维)的测试难度更大,测试的数量就更少。据报道目前已经测试过的生物大分子的独立三维结构不过 2000 余种^[1]。随着分子力学与分子动力学技术的兴起,特别是多拷贝随机分子动力学在药物创制中的应用,大量模拟分子的生物活性急需评价。这不仅需要相互作用能量的估

计,而且还需要定量结构与活性相关规律的建立。因此,从分子二维(2D)拓扑结构产生2D分子结构描述子,或先通过分子模拟技术产生三维(3D)结构再产生3D分子结构描述子,进而研究定量分子结构-活性/性质相关(quantitative structure-activity/property relationships,简称QSAR/QSPR)、完成目标靶分子与配体识别过程以及设计新型配体分子进而创制新型分子就显得特别重要和有实用意义^[2]。

大量事实表明,一个化合物的物理、化学或生物性质取决于分子中各个原子的三维构型。QSAR/QSPR研究近年来发展迅速,并已或正在对生物分子模拟、医农药、合成化学、药物创制、环境毒性评价等与生命科学、化学科学和环境科学等多个研究领域产生越来越深刻的影响。QSAR/QSPR技术已经成为研究物质理化性质与生物活性以寻求分子解释的所有学科领域中不可或缺的强有力工具^[3,4]。要建立QSAR/QSPR模型,必须首先对分子结构进行表征从而产生分子描述子(molecular descriptor,简称MD)。目前流行的MD包括二维拓扑描述子、能量描述子、量子力学描述子和三维分子场描述子等^[5]。另外还有用来描述基于分子在大量生物化学实验分析中的行为^[6]或基于目标分子对大量不同的蛋白质位点^[7]而产生的分子描述子。从描述分子结构的层次考虑,这些MD又可分为基于二维结构的所谓2D分子描述子和基于三维结构的3D分子描述子。从分子描述子个数考虑,MD又可分为单参数的标量描述子和多参数的“矢量”型描述子。

“矢量分子描述子”是本论文概括总结的一个概念,用来特指那些由多个等价的标量描述子共同表征一个分子结构的描述子集合。这里使用的“等价”是特别重要的,那些从不同方法产生或从不同参数库选择的编码不同结构特征信息的分子描述子集合不是矢量分子描述子。比如由编码不同结构信息的各种拓扑指数与量子力学描述子集合起来表征或描述一个分子的结构,这些MD集合就不是矢量分子描述子。我们提出的基于四原子类型的分子电

性距离矢量^[8,9]中包含的 10 个分子描述子,全部是基于二维分子结构的距离矩阵,以同样的计算公式从非氢原子的相对电性和原子之间的相对距离而得到的,是矢量分子描述子。又如各阶分子连接性指数的集合也是矢量分子描述子。由于分子描述子种类繁多,加上旧的方法不断改进,新的方法不断提出,要对所有这些分子描述子进行全面的分析讨论是不可能的,也是不可取的。本章仅以作者的见识和观点对属于“矢量”型的分子描述子作一些简单的讨论与分析。

1.2 矢量分子描述子研究的某些进展

1.2.1 分子连接性指数

各阶分子连接性指数系列(${}^m\chi$)集合,即零级、一级、二级、三级等分子连接性指数(${}^0\chi, {}^1\chi, {}^2\chi, {}^3\chi, \dots$)是一个典型的矢量分子描述子。分子连接性指数由 Randic^[10]提出并由 Kier 及其合作者扩展^[11~13]。业已表明分子连接性指数与多种物理化学性质存在良好的相关性,如分子极化度^[11,14]、水溶性与沸点^[14]、分配系数^[15]、范德华体积^[16]等。分子连接性指数也能够解释多种分子结构类型物质的生物化学性质包括大量环境领域中的多种应用^[17,18]。大量研究与应用表明,分子连接性是一个基于分子二维拓扑结构的最重要指数之一,可用于处理含环、多重键和含杂原子的分子体系,已成为应用最广泛的分子描述子之一^[19~22]。近年来,国内已有多本专著介绍分子连接性方法及其应用^[23,24],在此不再详述。

1.2.2 电拓扑状态指数

一个富含信息的分子描述子必须是计算快速、操作容易和易于解释的。正如在最近的几部著作^[25,26]和多篇研究论文^[27~31]中

所证实的那样，电拓扑状态及其相应描述子拥有上述列举的所有特征。电拓扑状态指数（electrotopological state index，简称 E-状态指数）已广泛应用于发展许多在原子水平^[25~28,31,32]和不同原子类型^[26,33~35]基础上的物理化学性质/生物活性相关模型。E-状态指数产生的结构信息，能明显地揭示分子结构特征与活性的关系。最近 Kier 他们又进一步地提出了氢原子水平的 E-状态指数^[25,26,31]更扩展了 E-状态作为一个强大的分子结构描述子的功能。各种 E-状态指数已经在许多 QSAR/QSPR 及其相关研究中进行了明确的定义。比如基于原子类型^[33]的 E-状态指数的构建一般包括以下几个步骤：首先根据各个非氢原子在分子中所处的局部化学环境对各原子所属类型进行划分，同时确定这些非氢原子价电子层所属主量子数、价层电子数、连接氢原子的数目、与其他原子形成的 σ 键和 π 键轨道数以及孤对电子数等；第二，计算各种原子类型的固有状态值（ I ）；第三，根据各原子在分子整体中与其他原子之间的拓扑距离，计算各个原子类型原子在其分子中固有状态值 I 的微扰项或增量（ ΔI ）；第四，计算各原子类型原子的固有状态值（ I ）及它与所有其他原子对它的微扰值（ ΔI ）之和（ S 值）；最后，将原子类型相同的各原子的 S 值相加即得该原子类型原子的 E-状态指数。分子中有多少种原子类型原子，就有多少个 E-状态指数，因此多原子类型的分子结构表征需要多个 E-状态指数，这些 E-状态指数集合即构成一个矢量分子描述子。事实已经证明，E-状态指数与各原子在分子中的固有状态有关，仅编码分子的二维拓扑结构信息，不需要分子的任何 3D 结构信息，是一个非常优秀的 2D 结构描述子。

1.2.3 分子指纹

分子指纹（fingerprint）是一类把分子结构特征表达成一个二进制位串格式的矢量分子描述子^[36~39]。分子指纹通常以一种二进制位串格式表征分子中某一个子结构碎片是否存在而获取有关分子

结构特征的信息^[40]。这些指纹常常都要经过一个散布程序^[41]以决定某个定义的结构特征在二进制位串中的位置。典型的指纹描述子之一称为 MACCS 译码(keys)。它首先定义 166 个或 960 个结构特征(每一个结构特征对应位串中的一个位点)组成所谓的“词典”这些结构特征构成 166 个或 960 个位点的一个二进制位串;然后根据某种结构特征在指定化合物中是否存在,决定该结构特征在词典中的相应位点位置的值为 1 或 0。由此可知,MACCS 指纹是一个严格意义上的结构“译码”,在位点与结构特征之间存在一一对应关系。另一个代表性的指纹描述子是阳光指纹(day-light fingerprint)。它不需要预先定义结构特征对应的位点“词典”,但是它都编码分子中原子以及路径等于 1 到 7 个键长的线性结构的有关信息。联合指纹(unity fingerprint)类似于阳光指纹,一个典型的 2D 联合指纹有 988 个位点,包括 928 个存在于化合物中的路径数位点和 60 个编码普通与稀有原子、一般原子类型、简单普通环的位点译码。BCI 指纹是相当独特的,它允许用户自己定义词典,可以编码 6 大类结构特征碎片,包括扩增原子(原子及与它们连接的键)原子序列(线性路径、默认 2 至 6 个原子)原子对(两个原子和它们之间的拓扑距离)环组分(原子序列和环周围的键)环混合(在每一个环周围环的连接数)稠环(单边稠环中非平面对称的立体构型)。

分子指纹已用于多种化学信息的处理系统中。例如, Martin 及其合作者使用阳光指纹,联合 $\log P$ 、拓扑指数和五个受体识别描述子,产生了一种表征大量化合物的方法^[42]。联合散布指纹已形成称为分子全息图的一个新的分子描述子的基础,并已应用于产生一组苯并二氮杂(benzodiazepine)与 GABA_A 受体成键活性的 QSAR 模型^[43]。与联合散布指纹不同的是,这个新的分子全息图的产生过程还包括一个碎片长度的选择问题,选择不同的碎片长度及其范围,会产生不同的分子全息图^[44]。这种全息 QSAR 方法^[45](hologram quantitative structure activity relations, 简称

HQSAR) 是一种不编码任何 3D 分子结构信息但具有与基于 3D 分子结构 QSAR 方法比如流行的 CoMFA 方法一样优秀的基于 2D 结构的 QSAR 方法, 是一种非常优秀的指纹分子描述子。

1.2.4 谱学描述子

实验谱学性质比如 NMR 化学位移、红外与拉曼光谱的伸缩频率等已广泛地用于化合物的表征与鉴定。用这些谱学性质作为分子描述子的主要问题是缺少预测能力。随着计算机技术的快速发展, 谱学性质可方便地通过从头算与半经验量子力学软件包进行计算, 虽然其可靠性随不同的计算方法有所不同, 但是都可以作为实验测定值的一种代换。最近提出的 EVA(normal coordinate eigenValues) 方法^[46,47], 应用基于分子坐标的量子力学计算得到的某些物理量如特征值与频率, 具有某种谱学数据的相似性。EVA 方法的步骤如下: 首先, 产生每一个化合物的三维分子结构坐标; 第二, 使用半经验计算程序包如 MOPAC 计算正常坐标频率; 第三, 将正常坐标特征值投影到一定宽度的频率标尺上, 比如红外光谱的 $0 \sim 4000 \text{ cm}^{-1}$ 上; 第四, 使用宽度为 σ 的高斯平滑函数于每一个振动频率上; 最后, 在每一个固定的频率间隔 (L) 采集强度值, 所有频率下的强度值即构成 EVA 矢量分子描述子。EVA 方法可以克服直接使用频率数据时每一个分子描述子的频率位置不同的困难, 因为正常模式的频率数量是随分子中所含原子数目的变化而变化的。虽然最近的报道表明 EVA 法可以用于控制分子的构象灵敏度^[48], 但其缺点是对每一个数据集都有两个参数 (L 和 σ) 需要调节与优化, 不同的参数 L 和 σ 将导致不同的 EVA 描述子。

EVA 描述子已成功用于正辛醇 - 水分配系数的预测^[46], 大量生物活性数据的估算^[49], 产生了一些可与优秀的三维 QSAR 方法具有可比性的模型, 但该方法不需要进行分子叠合。Benigni 及其合作者^[50]应用 EVA 描述子和三个其他描述子得到的模拟光谱

与实验红外光谱进行比较研究，得出了实验数据与计算光谱描述子一致的结论。Tuppurainen^[51]改进了 EVA 方法，提出了称为 EEVA 的有趣 EVA 方法，结果表明 EEVA 对大量生物活性数据集可产生相当好的模型。

1.2.5 CoMFA 与 GRID 描述子

CoMFA 原名 DYLOMMS(dynamic lattice oriented molecular modeling system,1987) 最早出现在 1979 年 4 月美国化学会会议上。CoMFA(comparative molecular field analysis 比较分子力场分析)自 1988 年^[52]公布以来，得到广泛应用^[53~55]，是目前三维 QSAR 中应用最广且最为成功的一种方法，并还在不断地发展完善之中。CoMFA 的基本步骤为：首先确定训练集中每一个分子合适的 3D 结构（药物分子的生物活性或药效构象）；第二，推导部分原子电荷以产生一个静电场；第三，使用合理的叠合方法叠合各分子；第四，在分子周围创建一个立方格子（一般应大于最大的分子）；第五，使用一个探针比如带一个单位正电荷的甲基计算相互作用能。这一步将产生一个基于 Lennard-Jones 6-12 势的立体相互作用能和一个基于 Coulombic 势的静电相互作用能，这些作用能即构成 CoMFA 描述子矢量；第六，以偏最小二乘（PLS, partial least-squares）方法^[56,57]提取相互作用能量与生物活性的定量关系，进而建立各分子生物活性与相互作用能的 QSAR 模型；最后，对检验集进行预测，可视化其结果，并将每一个分子的显示加上等高线。这只是一个总体简化的计算过程。大量的研究工作包括在上述各个步骤中，对其展开的争论（常常是相当激烈的）仍在继续。Norinder^[58]和朱杰^[59]对 CoMFA 方法学进展过程中，在化合物叠合策略（匹配规则）、新场引入、变量选择及统计学数据处理方法等四个方面进行了评述。在这四个问题中，也许最麻烦的是合适的叠合策略的选择。也许这需要利用某些与受体模型对接的算法^[60]。

CoMFA 最初是为小分子集而设计的。GRID^[61]和 CoMFA 非常类似,因为它们都要计算探针与小分子有时也有部分大分子之间的相互作用能。GRID 最初描述为一个计算程序,以确定重要生物大分子上的能量最低的成键位点。GRID 和 CoMFA 的主要差别是 GRID 使用多个探针确定相互作用能(大概是因为 GRID 起初的设计目标是探测生物大分子的成键位点),而 CoMFA 只使用一个探针。虽然 GRID 起初考察这些键合位点的愿望现在仍然是,但很明显它也能作为 CoMFA 类型方法中相互作用场的一种替换。除了探针如 CH_3 、 NH_2 、 NH_3^+ 和 O^- 等以外,主要的工作集中在 GRID 力场中氢键的参数化方面^[62,63]。Goodford 最近评述了 GRID 力场^[64]从该报告使用 GRID 的 CoMFA 类型方法的数量看,似乎多探针方法更普遍一些。流行的 GRID 版本提供了 56 个探针。其他类型的场也可添加到标准 CoMFA 方法中,第一个场是 Kellog 及其合作者引入疏水场的相互作用技术 HINT^[65-67]还有 HOMO^[68]和 LUMO^[69]场等等。毫无疑问,新的相互作用场的发展将继续推进 CoMFA 和相关技术的进步。

在 GRID 和 CoMFA 之间的其他主要差别是 PLS 的应用。PLS 非 GRID 方法所必需但属 CoMFA 整体的一部分。GRID 最初设计时,只需要简单计算相互作用场,然后使用分子图形技术进行可视化,只是在应用于 CoMFA 类型的方法中时才使用 PLS 来建立回归模型。

1.2.6 与 CoMFA 相关的描述子

两个与 CoMFA 相关的技术是 CoMSIA (comparative molecular similarity index analysis, 比较分子相似性指数分析)^[70]和 CoMMA (comparative molecular moment analysis, 比较分子动量矩分析)^[71]。CoMSIA 利用一个共同探针,计算其与整个规则空间格子上每一个分子之间的相似性量度。相似性的距离量度通过高斯函数计算,据称这可避免原子位置附近的奇异值,这也不同于在标

准 CoMFA 方法中使用的 Lennard - Jones 6 - 12 势和 Coulombic 势。在应用五种不同性质力场对嗜热菌蛋白酶抑制剂的分析中发现 CoMSIA 方法结果比标准 CoMFA 更易解释^[72]。CoMMA 方法则利用分子动量矩以及由化合物的二极矩和四极矩推导的相关性质^[73]。该技术的优点之一也是不需要叠合步骤，因为分子叠合是大多数三维 QSAR 方法中都要碰到的共同问题，关于这个问题，最近有一篇报告作了较详细的讨论^[74]。另一个报道的不需要叠合步骤的方法是基于测量的与模拟的光谱信息的一种技术，即 CoSA (comparative spectra analysis 比较光谱分析)^[75]。CoSA 已用于众所周知的标准甾族化合物的 QSAR 研究，得到可与 CoMFA 相比或更好的结果。有趣的是，将光谱描述子与 CoMFA 场组合起来在大多数情况下具有更强大的性能。

最后要提到的一个方法是 HASL^[76,77] (hypothetical active site lattice 假想活性位点网格)。该技术采用训练集中各个分子性质的格子。从某种意义上说，操作这个格子，是在与 CoMFA 相对的方向进行的。在构建 CoMFA 格子时，产生一个格子包围所有的分子，但那些非常接近或在该分子 van der Waals 半径的格点将被丢弃。HASL 也使用格点，但选为网格的是那些处于 van der Waals 半径以内的那些点。根据最近原子的性质，给这些网格点赋予性质值 (HASL 型值)，然后这些网络对每一个化合物进行操作而形成描述子。该方法已有成功的应用^[77]。此外，一组抗疟疾 artemisinin 模拟物的 QSAR 研究结果表明，HASL 方法具有与 CoMFA 完全可比的质量^[78]。同时这个报告得出结论说 HASL 与 CoMFA 表达了两种相互补充的方法。

1.2.7 分子相似性描述子

Ghuloum^[79]等提出了一种表征分子表面性质的称为 Hashkey 的分子数字表征方法。Hashkey 是一个能捕获一个分子有关表面性质信息的固定维数的通用而简洁的矢量分子描述子方法。具有

相似 Hashkey 矢量的一组分子，其表面性质具有相似性。或者说，具有相同表面性质的分子具有相同的 Hashkey，而与该表面之内所包含的原子骨架无关。一个给定分子 M 其 N 维 Hashkey ($H_1, H_2, \dots, H_N$) 可通过计算相对于一组基分子 (N 个 $B_1, B_2, \dots, B_N$) 的表面相似性^[80~82]获得。这些基分子应该是固定构象和低能量的。给定分子 M 被柔性地叠合到这个基分子集中的每一个基分子并且使分子 M 与 B_i 的表面相似性达到最大化，即达到最好匹配，从而产生表面相似性值 H_i 。Hashkey 总是对一个分子整体进行计算。为了考察其方法在药理学性质方面的预测性能，他们应用综合临床医药化学数据库 (CMC 数据库) 中的数据创建了能够与流行方法一样准确地预测正辛醇-水分配系数的模型。另外他们还利用结构多样性药物代谢数据和分子 Hashkey 预测了人肠内吸收活性。

1.2.8 WHIM 描述子

WHIM 描述子 (weighted holistic invariant molecular descriptor, 加权整体不变量分子描述子) 起源于意大利米兰的 Todeschini 研究组^[83~85]。这个迷人的三维 QSAR 方法是基于原子坐标的主成分分析结果，据称是一个对旋转和平移的不变量，因而不需要叠合程序。原始 WHIM 过程包括以下几步：首先，将一个加权方案应用到分子坐标；第二，坐标中心化后计算加权协方差矩阵；第三，该协方差矩阵作为主成分分析的基础；最后，原子的得分矩阵被用来提取统计参数，进而推导 WHIM 描述子。因为坐标被中心化，所以产生的描述子对平移是一个不变量；主成分分析步骤给出了唯一对旋转为不变量的解答。所以，一个数据集的 WHIM 描述子是不受数据集中化合物的相对定位的影响的。加权方案由单位加权 (实际是不加权)、原子质量、van der Waals 体积、Mulliken 原子电负性、原子极化度和 Kier 与 Hall 的电拓扑指数等组成。对于每

一个加权方案，都定义了两种 WHIM 描述子，即直接的和非直接的。这些参数可从各种各样的主成分分析结果中计算出来，例如，特征值 λ_1 、 λ_2 、 λ_3 等用作“大小”描述子，特征值的比例用作形状描述子。一共有 66 个直接的 WHIM 描述子（如果是平面分子则为 48 个）和 37 个非直接的 WHIM 描述子^[86]。WHIM 描述子的最初应用表明，它们可以用于描述大量的物理化学性质比如沸点、溶解度、摩尔折光率、极化度等^[84,85,87,88]。这是不奇怪的，因为它们也表明了能成功地建立生物响应模型，范围从毒性^[89]到肾上腺素拮抗剂^[88]。因为 WHIM 方法基于一组 3D 坐标，它可以应用于分子的表征。最近报道表明，WHIM 已扩展到考察分子表面^[90]和基于格子的相互作用场的处理^[91]。不用怀疑，WHIM 描述子含有有用的化学信息。它们具有优势，可以对任何化合物进行计算，给出一个合理的可用的 3D 结构，而且它与其他 3D QSAR 方法相比不需要叠合策略，因而具有非常明显的好处。

1.2.9 小结

分子结构表征或分子描述子研究是一个相当广阔的领域。目前已出现大量优秀的 2D 与 3D 分子描述子，它们都具有良好的模型估计与预测能力，都已广泛地应用于分子模拟的各个领域中。虽然从理论上讲，3D 方法更接近实际，应该更加优越，但就目前情况看，由于 3D 方法需要首先对分子结构进行三维结构搜索并依据能量最低原理进行优化以获得优势构象或寻求生物活性构象以及实施叠合策略，因此要构建有效的基于三维分子结构的 3D QSAR 模型仍是相当困难和很耗时的。正因为如此，在发展 3D QSAR 技术的同时，许多优秀的 2D QSAR 方法仍在不断地改进，新的 2D QSAR 方法也在不断地提出。事实证明，许多优秀的 2D 方法具有与复杂的 3D 方法一样的模型稳定性和可预测性^[92]。因此，在生物信息学与化学信息学逐渐成熟的今天，优秀 2D 与 3D 方法具有同样的生命力，正在各自发挥自己强有力的功能。

1.3 本文的研究内容

基于有机物分子的二维拓扑结构,按照预定义的原子类型与原子属性分类方案,考虑非氢原子相对电性或电拓扑状态指数、相对距离或拓扑距离,创建了一种新型矢量分子描述子,即三种形式^[8,9,93,94]的分子电性距离矢量(molecular electronegativity distance vector 简称 MEDV)。将这些 MEDV 矢量分子描述子应用于包括药物与环境污染物的有机物分子结构的表征及其理化性质与生物活性的 QSAR/QSPR 研究以说明 MEDV 方法的适用性和应用范围。本文的有机物分子电性距离矢量表征及其应用的主要研究内容包括 MEDV 描述子的建立、研究方法及其在各种有机物物理、化学与生物学性质的估计与预测中的应用等。MEDV 描述子及其应用的基本框架如图 1-1 所示。全文将分 6 章进行介绍与讨论。

第一章的绪论简要介绍分子结构表征的意义、分子描述子的分类以及矢量分子描述子的某些进展等。

第二章详细说明三个分子电性距离矢量(MEDV 描述子的构造方法。首先介绍 MEDV 方法中所用到的“原子类型”、“原子属性”、“相对电性”与“相对距离”等基本概念,进而从简单到复杂,由特殊到一般,叙述我们提出的三种类型分子电性距离矢量描述子的计算方法。这三种 MEDV 描述子包括:基于四种原子类型的分子电性距离矢量 MEDV-4(简称 MEDV)、分子全息距离矢量 MHDV 和基于十三种原子类型的分子电性距离矢量 MEDV-13。

第三章讨论矢量分子描述子的研究方法。首先讨论如何评价矢量分子描述子对分子结构的分辨率;进而讨论 QSAR/QSPR 研究中变量优化选择与模型化方法以及以交互检验为目标函数的模型质量评价方法体系。主要介绍以交互检验质量为目标函数的最佳子集回归(BSR)、主成分回归(PCR)和偏最小二乘回归(PLSR)

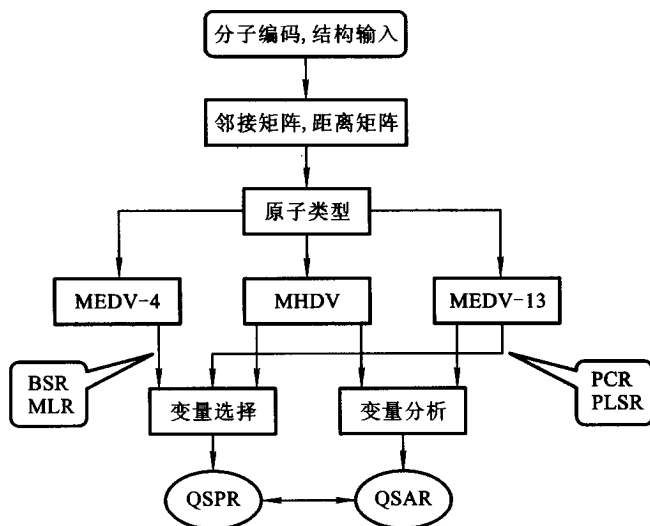


图 1-1 分子结构表征及其应用研究流程简图

等三种方法。最佳子集回归方法主要用于描述子变量较少的分子体系比如在 MEDV-4 方法建立 QSAR/QSPR 模型时, 而 PCR 和 PLSR 方法则主要用于变量数较多的大分子体系比如在 MHDV 和 MEDV-13 方法中。

第四章重点说明 MEDV-4 矢量分子描述子对有机分子结构的表征及其在 QSPR 研究中的应用。在这一章我们应用 MEDV-4 与最佳子集回归方法建立了烷烃、烯烃、炔烃、多环芳烃等多种有机物的沸点、密度、临界压力、色谱保留指数等十多种物理化学性质与热力学函数的 QSPR 模型。

第五章将 MHDV 和 MEDV-13 矢量分子描述子应用于几类典型生物活性物质的活性估计与预测。介绍了 MHDV 矢量描述子对两组二肽药物分子生物活性及一组取代苯化合物的生物可降解性的估计与预测, 通过与有关文献结果的比较, 说明 MHDV 方

法的有效性。给出了 MEDV - 13 方法在建立“标准”甾族化合物亲和能模型与肠促胰酶肽 - A 受体拮抗剂的生物活性 QSAR 模型中的应用。

第六章将 MEDV - 13 矢量分子描述子应用于几类选择性抑制环加氧酶 - 2 (cyclooxygenase - 2, COX - 2) 非甾类抗炎药物分子结构的表征及其抑制活性的 QSAR 研究。研制高效、低毒、副作用小的非甾类抗炎药物近年受到广泛重视。选择性抑制 COX - 2 酶的药物可以降低或消除现有非甾类抗炎药物由于不加选择地抑制两种环加氧酶而引起的胃肠道刺激、损伤和溃疡等毒副作用，因而具有重要意义。

参 考 文 献

- [1] Thornton J M, Todd A E, Milburn D, et al. From structure to function: approaches and limitations. *Nature Struct. Biol.*, 2000, 7 (Suppl.): 991 ~ 994
- [2] Walters W P, Stahl M T, Murcko M A. Virtual Screening: an overview. *Drug Discovery Today*, 1998, 3: 160 ~ 178
- [3] Katritzky A R, Maran U, Lobanov V S, Karelson M. Perspective: structurally diverse quantitative structure - property relationship correlations of technologically relevant physical properties. *J. Chem. Inf. Comput. Sci.*, 2000, 40 (1): 1 ~ 18
- [4] Robinson D D, Winn P J, Lyne P D, Richards W G. Self - organizing molecular field analysis: a tool for structure - activity studies. *J. Med. Chem.*, 1999, 42: 573 ~ 583
- [5] Patterson D E, Cramer R D, Ferguson A M, Clark R D, Weinberger L E. Neighborhood behavior: a useful concept for validation of “molecular diversity” descriptors. *J. Med. Chem.*, 1996, 39: 3049 ~ 3059
- [6] Kauvar L M, Higgins D L, Villar H O, et al. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.*, 1995, 2: 107 ~ 108
- [7] Briem H, Kuntz I D. Molecular similarity based on dock-generated finger-

prints. J. Med. Chem. ,1996,39:3401~3408

- [8] Liu S S, Cao C Z, Li Z L. Approach to estimation and prediction for normal boiling point (NBP) of alkanes based on a novel molecular distance-edge (MDE) vector, λ . J. Chem. Inf. Comput. Sci. ,1998,38(3):387~394
- [9] 刘树深 刘堰 李志良 蔡绍哲. 一个新的分子电性距离矢量 (MEDV). 化学学报. 2000,58(11):1353~1357
- [10] Randic M. On characterization of molecular branching. J. Am. Chem. Soc. ,1975,97:6609~6615
- [11] Kier L B, Hall L H, Murray W J, Randic M. Molecular connectivity 1: relationship to nonspecific local anesthesia. J. Pharm. Sci. ,1975,64:1971~1974
- [12] Kier L B, Murray W J, Hall L H. Molecular connectivity 4: relationship to biological activity. J. Med. Chem. ,1975,18:1272~1274
- [13] Kier L B, Hall L H. Molecular connectivity analysis of structure influencing chromatographic retention indexes. J. Pharm. Sci. ,1979,68:120~122
- [14] Hall L H, Kier L B, Murray W J. Molecular connectivity 2: relationship to water solubility and boiling points. J. Pharm. Sci. ,1975,64:1974~1977
- [15] Murray W J, Hall L H, Kier L B. Molecular connectivity 3: partition coefficients. J. Pharm. Sci. ,1975,64:1978~1981
- [16] Hall L H, Kier L B. The relation of molecular connectivity to molecular volume and biological activity. Eur. J. Med. Chem. ,1981,16:399~407
- [17] Nirmalakhandan N N, Speece R E. Structure-activity relationships. Quantitative techniques for predicting the behavior of chemicals in the ecosystem. Environ. Sci. Technol. ,1988,22:606~615
- [18] Okey R W, Stensel H D. A QSAR development procedure for aromatic xenobiotic degradation by unacclimated bacteria. Water Environ. Res. ,1993,65:772~780
- [19] Cash G G, Breen J J. Correlation of graph-theoretical parameters with biological activity. J. Chem. Inf. Comput. Sci. ,1993,33:275~279
- [20] Llorente B, Rivero N, Carrasco R, et al. A QSAR study of quinolones based on electrotopological state index for atoms. Quant. Struct.-Act. Relat. ,1994,13:419~425

- [21] Osmialowski K, Kaliszan R. Studies of performance of graph theoretical indexes in QSAR analysis. *Quant. Struct. -Act. Relat.*[†], 1991, 10:125~134
- [22] Shapiro S, Guggenheim B. Inhibition of oral bacteria by phenolic compounds. Part 1. QSAR analysis using molecular connectivity. *Quant. Struct. -Act. Relat.*, 1998, 17:327~337
- [23] 王连生, 支正良. 分子连接性与分子结构 - 活性, 北京: 中国环境科学出版社, 1992
- [24] 王连生, 韩朔睽. 分子结构、性质与活性, 北京: 化学工业出版社, 1997
- [25] Kier L B, Hall L H. The electrotopological state: structure modeling for QSAR and database analysis. In *Topological indices and related descriptors in QSAR and QSPR*. Reading, U. K. :Gordon and Breach, 1999
- [26] Hall L H, Kier L B. Molecular structure description: the electrotopological state. San Diego:Academic Press, 1999
- [27] AbouShaaban R R A, AlKhamees H A, AbouAuda H S, et al. Atom level electrotopological state indexes in QSAR: designing and testing antithyroid agents. *Pharmaceut. Res.*, 1996, 13(1):129~136
- [28] Hall L H, Mohnhey B, Kier L B. Comparison of electrotopological state indexes with molecular orbital parameters: inhibition of MAO by hydrazides. *Quant. Struct. -Act. Relat.*, 1993, 12:44~48
- [29] Kier L B, Hall L H. An electrotopological state index for atoms in molecules. *Pharm. Res.*, 1990, 7:801~807
- [30] Kier L B, Hall L H. An index of electrotopological state for atoms in molecules. *J. Math. Chem.*, 1991, 7:229~241
- [31] Kier L B, Hall L H. Inhibition of salicylamide binding: a electrotopological state analysis. *Med. Chem. Res.*, 1992, 2:497~502
- [32] Hall L H, Mohnhey B K, Kier L B. The electrotopological state: structure information at the atomic level for molecular graphs. *J. Chem. Inf. Comput. Sci.*, 1991, 31:76~82
- [33] Hall L H, Kier L B. Electrotopological state indexes for atom types: a novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.*, 1995, 35:1039~1045
- [34] Hall L H, Story C T. Boiling point and critical temperature of a heteroge-

neous data set: QSAR with atom type electrotopological state indices using artificial neural networks. *J. Chem. Inf. Comput. Sci.*, 1996, 36: 1004 ~ 1014

- [35] Hall L H, Story C T. Boiling point of a set of alkanes, alcohols and chloroalkanes: QSAR with atom type electrotopological state indices using artificial neural networks. *SAR QSAR Environ. Res.*, 1997, 6: 139 ~ 161
- [36] Flower D R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.*, 1998, 38: 379 ~ 386
- [37] McGregor M J, Pallai P V. Clustering of large databases of compounds: Using MDL "Keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.*, 1997, 37: 1 ~ 9
- [38] Wild D J, Blankley C J. Comparison of 2D fingerprint types and hierarchy level selection methods for structural grouping using Ward's clustering. *J. Chem. Inf. Comput. Sci.*, 2000, 40: 155 ~ 162
- [39] Xue L, Godden J W, Bajorath J. Database searching for compounds with similar biological activities using short binary bit string representations of molecules. *J. Chem. Inf. Comput. Sci.*, 1999, 39: 881 ~ 886
- [40] Willett P. Computational tools for the analysis of molecular diversity. *Perspect. Drug Discov. Des.*, 1997, 718: 1 ~ 11
- [41] Ihlenfeldt W D, Gasteiger J. HASH codes for the identification and classification of molecular structure elements. *J. Comput. Chem.*, 1994, 15: 793 ~ 813
- [42] Martin E J, Blaney J M, Siani M S, et al. Measuring diversity: experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.*, 1995, 38: 1431 ~ 1436
- [43] Winkler D A, Burden F R. Holographic QSAR of benzodiazepines. *Quant. Struct. - Act. Relat.*, 1998, 17: 224 ~ 231
- [44] Seel M, Turner D B, Willet P. Effect of parameter variations on the effectiveness of HQSAR analysis. *Quant. Struct. - Act. Relat.*, 1999, 18: 245 ~ 252
- [45] Tong W, Lewis D R, Perkins R, Chen Y, Welsh W J, Goddette D W, Heritage T W, Sheehan D M. Evaluation of quantitative structure-activity rela-

- tionship methods for large – scale prediction of chemicals binding to the estrogen receptor. *J. Chem. Inf. Comput. Sci.* , 1998, 38:669 ~ 677
- [46] Ferguson A M, Heritage T, Jonathon P, et al. EVA: a new theoretically based molecular descriptor for use in QSAR/QSPR analysis. *J. Comput. -Aided Mol. Des.* , 1997, 11:143 ~ 147
- [47] Ginn C M R, Turner D B, Willett P, et al. Similarity searching in files of three – dimensional chemical structures: Evaluation of the EVA descriptor and combination of rankings using data fusion. *J. Chem. Inf. Comput. Sci.* , 1997, 37:23 ~ 27
- [48] Turner D B, Willett P, Ferguson A M, et al. Evaluation of a novel molecular vibration-based descriptor(EVA)for QSAR studies 2: Model validation using a benchmark steroid dataset. *J. Comput. -Aided Mol. Des.* , 1999, 13: 271 ~ 296
- [49] Turner D B, Willett P, Ferguson A M, et al. Evaluation of a novel infrared range vibration – based descriptor(EVA)for QSAR studies. 1. General application. *J. Comput. -Aided Mol. Des.* , 1997, 11:409 ~ 422
- [50] Benigni R, Passerini L, Livingstone D J, et al. Infrared spectra information and their correlation with QSAR descriptors. *J. Chem. Inf. Comput. Sci.* , 1999, 39:558 ~ 562
- [51] Tuppurainen K. EEVA(electronic eigenvalue): a new QSAR/QSPR descriptor for electronic substituent effects based on molecular orbital energies. *SAR QSAR Environ. Res.* , 1999, 10:39 ~ 46
- [52] Cramer R D III , Patterson D E, Bunce J D. Comparative molecular field analysis(ComFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* , 1988, 110:5959 ~ 5967
- [53] DePriest S A. 3D-QSAR of angiotensin – converting enzyme and thermolysin inhibitors: a comparison of ComFA models based on deduced and experimentally determined active site geometries. *J. Am. Chem. Soc.* , 1993, 115: 5372 ~ 5384
- [54] Horwitz J P. Comparative molecular field analysis of in vitro growth inhibition of L1210 and HCT-8 cells by some pyrazoloacridines. *J. Med. Chem.* , 1993, 36:3511 ~ 3516

- [55] Klebe G, Abraham U. On the prediction of binding properties of drug molecules by comparative molecular field analysis. *J. Med. Chem.*, 1993, 36: 70 ~ 80
- [56] Baroni M, Clementi S, Cruciani G, et al. Predictive ability of regression models. Part 2. Selection of the best predictive PLS model. *J. Chemom.*, 1992, 6: 347 ~ 356
- [57] Cruciani G, Watson K A. Comparative molecular field analysis using GRID force field and GOLPE variable selection methods in a study of inhibitors of glycogen. *J. Med. Chem.*, 1994, 37: 2589 ~ 2601
- [58] Norinder U. Recent progress in CoMFA methodology and related techniques. *Perspect. Drug Discov. Des.*, 1998, 12/13/14: 25 ~ 39
- [59] 朱杰 盛春泉 张万年. 比较分子力场分析法 (CoMFA) 的研究新进展. *化学进展* 2000, 12: 203 ~ 207
- [60] Jones G, Willet P. Docking small-molecule ligands into active sites. *Curr. Opin. Biotechnol.*, 1995, 6: 652 ~ 656
- [61] Goodford P J. A computational procedure for determining energetically favorable binding sites on biologically important molecules. *J. Med. Chem.*, 1985, 28: 849 ~ 857
- [62] Wade R C, Clark K, Goodford P J. Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 1. Ligand probe groups with the ability to form two hydrogen bonds. *J. Med. Chem.*, 1993, 36: 140 ~ 147
- [63] Wade R C, Clark K, Goodford P J. Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 2. Ligand probe groups with the ability to form more than two hydrogen bonds. *J. Med. Chem.*, 1993, 36: 148 ~ 156
- [64] Goodford P J. Multivariate characterization of molecules for QSAR analysis. *J. Chemom.*, 1996, 10: 107 ~ 117
- [65] Kellogg G E, Semus S F, Abraham D J. HINT: a new method of empirical hydrophobic field calculations for CoMFA. *J. Comput.-Aid. Mol. Des.*, 1991, 5: 545 ~ 552
- [66] Kellogg G E, Joshi G S, Abraham D J. New tools for modeling and under-

- standing hydrophobicity and hydrophobic interactions. *Med. Chem. Res.* , 1991, 1:444 ~ 453
- [67] Kellogg G E, Abraham R J. KEY, LOCK, LOCKSMITH: complementary hydropathic map predictions of drug structure from a known receptor-receptor structure from known drugs. *J. Mol. Graph.* , 1992, 10:212 ~ 217
- [68] Waller C L, Marshall G R. 3D-QSAR three dimensional quantitative structure-activity relationship of angiotensin-converting enzyme incorporating molecular orbital fields and desolvation free energies based on active-analog and complementary-receptor-field alignment rules. *J. Med. Chem.* , 1993, 36:2390 ~ 2403
- [69] Poso A, Tuppurainen K, Gynther J. Modeling of molecular mutagenicity with comparative molecular field analysis(ComFA). Structural and electronic properties of MX compounds related to TA100 mutagenicity. *J. Mol. Struct. (THEOCHEM)* , 1994, 304:255 ~ 260
- [70] Klebe G, Abraham U, Mietzner T. Molecular similarity indices in a comparative analysis(ComSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* , 1994, 37:4130 ~ 4146
- [71] Silverman B D, Platt D E. Comparative molecular moment analysis(ComMA): 3D-QSAR without molecular superposition. *J. Med. Chem.* , 1996, 39:2129 ~ 2140
- [72] Klebe G, Abraham U. Comparative molecular similarity index analysis (ComSIA) to study hydrogen - bonding properties and to score combinatorial libraries. *J. Comput. -Aided Mol. Des.* , 1999, 13:1 ~ 10
- [73] Silverman B D, Platt D E, Pitman M C, Rigotsos I. Comparative molecular moment analysis. *Perspect. Drug Discov. Des.* , 1998, 12/13/14:183 ~ 196
- [74] Silverman B D, Pitman M C, Platt D E, Rigotsos I. Molecular moment similarity between clozapine and substituted [(4-phenylpiperazinyl)-methyl] benzamides; selective dopamine D4 agonists. *J. Comput. -Aided Mol. Des.* , 1998, 12:525 ~ 532
- [75] Bursi R, Dao T, van Wijk T, et al. Comparative spectra analysis(CoSA): spectra as three - dimensional molecular descriptors for the prediction of biological activities. *J. Chem. Inf. Comput. Sci.* , 1999, 39:861 ~ 867

- [76] Doweyko A M. The hypothetical active site lattice. An approach to modeling active sites from data on inhibitor molecules. *J. Med. Chem.* , 1988, 31: 1396 ~ 140
- [77] Doweyko A M. Pharmacophores from binding data. *J. Med. Chem.* , 1994, 37: 1769 ~ 1778
- [78] Woolfrey J R, Avery M A, Doweyko A M. Comparison of 3D quantitative structure – activity relationship methods: analysis of the in vitro antimalarial activity of 154 artemisinin analogs by hypothetical active-site lattice and comparative molecular field analysis. *J. Comput.-Aided Mol. Des.* , 1998, 12: 165 ~ 181
- [79] Ghuloum A M, Sage C R, Jain A N. Molecular Hashkeys: a novel method for molecular characterization and its application for predicting important pharmaceutical properties of molecules. *J. Med. Chem.* , 1999, 42: 1739 ~ 1748
- [80] Blandan P. A rapid method for comparing and matching the spherical parameter surfaces of molecules and other irregular objects. *J. Mol. Graph.* , 1989, 7: 130 ~ 137
- [81] Blaney F, Finn P, Phippen R, Wyatt R. Molecular surface comparison: application to drug design. *J. Mol. Graph.* , 1993, 11: 98 ~ 105
- [82] Jain A N, Harris N L, Park J Y. Quantitative binding site model generation: compass applied to multiple chemotypes targeting the 5Hta receptor. *J. Med. Chem.* , 1995, 38: 1295 ~ 1307
- [83] Todeschini R, Lasagni M, Marengo E. New molecular descriptors for 2D and 3D structures. *Theory. J. Chemom.* , 1994, 8: 263 ~ 272
- [84] Todeschini R, Gramatice P, Provenzani R. Weighted holistic invariant molecular descriptors. Part 2. Theory development and applications on modeling physicochemical properties of polycyclic aromatic hydrocarbons. *Chemom. Intell. Lab. Syst.* , 1995, 27: 221 ~ 229
- [85] Todeschini R, Gramatice P. 3D – modeling and prediction by WHIM descriptors. Part 5. Theory development and chemical meaning of WHIM descriptors. *Quant. Struct. -Act. Relat.* , 1997, 16: 113 ~ 119
- [86] Todeschini R, Gramatice P. The WHIM theory: new 3D molecular descriptors for QSAR in environmental modeling. *SAR QSAR Environ. Res.* ,

1997,7:89~115

- [87] Chiorboli C, Gramatice P, Piazza R, et al. 3D – modeling and prediction by WHIM descriptors. Part 7. Physico-chemical properties of haloaromatics: comparison between WHIM and topological descriptors. SAR QSAR Environ. Res. ,1997,7:133~150
- [88] Todeschini R, Gramatice P. 3D – modeling and prediction by WHIM descriptors. Part 6. Application of WHIM descriptors in QSAR studies. Quant. Struct. -Act. Relat. ,1997,16:120~125
- [89] Todeschini R, Vighi M, Provenzani R, et al. Modeling and prediction by using WHIM descriptors in QSAR studies: toxicity of heterogeneous chemicals on *Daphnia magna*. Chemosphere,1996,32:1527~1545
- [90] Bravi G, Gancia E, Mascagni P, Pegna M, Todeschini R, Zaliani A. MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: a comparative 3D QSAR study in a series of steroids. J. Comput. -Aided Mol. Des. ,1997,11:79~92
- [91] Todeschini R, Moro G, Boggia R, et al. Modeling and prediction of molecular properties. Theory of grid – weighted holistic invariant molecular (G-WHIM) descriptors. Chemom. Intell. Lab. Syst. ,1997,36:65~73
- [92] Brown R D, Martin Y C. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. J. Chem. Inf. Comput. Sci. , 1997,37:1~9
- [93] Liu S S, Yin C S, Cai S X, Li Z L. A novel MHDV descriptor for dipeptide QSAR studies. J. Chin. Chem. Soc. ,2001,48(2):253~260
- [94] Liu S S, Yin C S, Cai S X, Li Z L. QSAR study of steroid benchmark and dipeptides based on MEDV-13. J. Chem. Inf. Comput. Sci. ,2001,41(2):321~329

第2章

建立分子电性距离矢量

大多数有机化合物分子如药物和有机污染物分子都是由具有不同电（负）性和共价半径的各个原子通过化学键主要是共价键构成的。要对药物与有机污染物进行定量结构 - 性质 / 活性关系 (QSPR/QSAR 或者定量结构 - 毒性关系 QSTR) 研究 进而进行分子结构改造或设计新型药物分子，最后实施对这些物质分子的合成与利用，必须对这些分子结构进行定量表征，获得这些物质的分子描述子 (molecular descriptor)。“分子电性距离矢量”(molecular electronegativity distance vector 简称 MEDV) 是我们根据分子二维拓扑结构，利用分子环境中各原子的电性特征与原子间距离构建的一类新型矢量分子描述子，共有三种类型。第一种是基于 4 种原子类型，应用各非氢原子相对电性与相对距离信息构建的“分子电性距离矢量”(MEDV - 4)，适用于描述分子结构较为简单，杂原子种类较少的分子体系；第二种是基于 13 种原子类型，充分考察各种不同化学键的相对键长构建的“分子全息距离矢量”(MHDV)，适用于描述分子中化学键类型较多，各化合物原子类型比较一致的分子体系；第三种是基于 13 种原子类型，充分考虑非氢原子在不同分子结构中的电拓扑性质构建的“分子电性距离

矢量 χ (MEDV-13)，是一种通用型分子电性距离矢量，适用于各种有机化合物结构的表征。通过对大量有机分子体系分子结构的分子电性距离矢量表征与定量结构-性质/活性相关研究，表明该矢量描述子对各种化合物结构具有良好的分辨率，与分子的各种物理化学性质或生物活性具有良好的相关性。本章将详细讨论该矢量描述子的基本概念、类型、构建方法以及算法实现等基本理论。

2.1 基于四原子类型的 分子电性距离矢量 (MEDV-4)

基于4种原子类型的“分子电性距离矢量”是基于分子二维拓扑结构，以各种非氢原子的相对电负性以及这些原子之间的相对距离为主要分子结构特征而提出的一个矢量分子描述子，其英文名称定义为：molecular electronegativity distance vector based on 4 atomic types，简称 MEDV-4。

2.1.1 MEDV-4 中的几个基本概念

要建立分子电性距离矢量 MEDV-4，首先要定义各种分子结构中不同原子主要是非氢原子的原子类型 (atomic type) 和原子属性 (atomic attribute)，同时还要给出电性与距离的量度标准。因此，先讨论分子电性距离矢量所涉及的一些基本问题。

1. 原子类型 (atomic type)

众所周知，对于大多数有机化合物分子，主要是由 C, H, O, N, S, P, F, Cl, Br, I 等少数几种非金属元素原子通过共价键以各种不同的连接方式键合而成的。这些有机分子都是以碳原子为骨架的。由于 H 原子的特殊性，人们在研究这些化合物时常常仅考虑由各个重原子或称非氢原子构成的有机分子骨架。或者说，这些有机物分子可以简单地以非氢原子骨架结构来表示，因为由这个

骨架可方便地得到完整的分子结构。为了说明这些非氢原子在有机分子中所处的不同化学环境或不同拓扑结构状态，我们首先定义了原子类型概念^[1~3]。

有机分子中某非氢原子的原子类型(atomic type)定义为该非氢原子在分子中与其他非氢原子通过共价键相互连接的非氢原子个数。如果某非氢原子与 k 个非氢原子以共价键相连，则该非氢原子就属于第 k 原子类型。这个共价键可以是单键、双键、三键，也可以是其他类型化学键。在这里我们指定原子类型用小写斜体字母 k 或 l 表示。比如某非氢原子属于第 1 种原子类型，就称它为 $k=1$ 或 $l=1$ 。对于已经发现与合成的绝大多数有机化合物（如药物或有机污染物）分子中各种非氢原子最多只有 4 种原子类型，也即这些分子中的各个非氢原子最多只可能与另外 4 个非氢原子直接相连。表 2-1 给出二甲基氨基苯基甲烷分子（分子结构及隐氢结构如图 2-1）中 10 个非氢原子的原子类型。

表 2-1 苯基氨基甲烷分子中 10 个非氢原子的原子类型和原子属性

原子编号	1	2	3	4	5	6	7	8	9	10
原子类型	2	2	2	2	2	3	4	1	1	1
原子属性	CC	CC	CC	CC	CC	CC	C1	N1	C1	C1

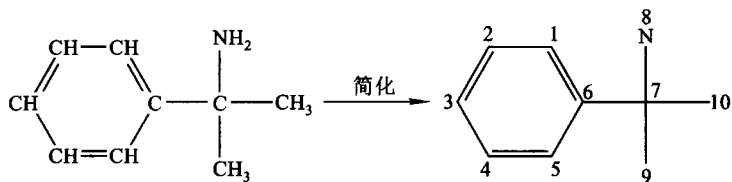


图 2-1 二甲基氨基苯基甲烷的结构与隐氢图

2. 原子属性 (atomic attribute)

为了反映非氢原子在分子中是通过什么样的共价键与其他非

氢原子相连的，定义了原子属性 (atomic attribute) 的概念。原子属性是一个用于表征处于化学环境中的非氢原子所属元素原子特征以及在分子中所属化学键特性的一种符号，采用一个双字符的字符串表示 (表 2-2 列出了有机物分子中常见的原子属性)。原子属性定义规定：第一个字符表示该非氢原子所属化学元素原子类型属性，一般用化学元素符号表示，但氯与溴分别以 L, B 表示；第二个字符表示与其他非氢原子形成共价键的类型属性，即什么样的共价键。例如与第一个表征元素原子类型的字符重复时表示形成的化学键是共轭双键，而数字 1, 2, 3 分别表示形成单键、双键及三键等。形成多个共价键时，以共价键的键级高低为优先次序。比如以“三键→共轭双键或双键→单键”的次序定义原子属性。由此定义，可以推断表 2-1 中所示二甲基苯基氨基甲烷分子中的 10 个非氢原子的原子属性按编号从小到大的顺序分别表示即为：CC, CC, CC, CC, CC, CC, C1, N1, C1 和 C1 (也可参见表 2-1)。

表 2-2 有机及药物分子中常见的原子属性

原子属性	说明	原子属性	说明	原子属性	说明
C1	sp ³ 杂化碳	NN	共轭体系中含氮	S1	单键中硫
C2	sp ² 杂化碳	P1	单键中磷	S2	双键中硫
C3	sp 杂化碳	P2	双键中磷	SS	共轭体系中含硫
CC	共轭体系中含碳	PP	共轭体系中含磷	F1	单键中氟
N1	单键中氮	O1	单键中氧	L1	单键中氯
N2	双键中氮	O2	双键中氧	B1	单键中溴
N3	三键中氮	OO	共轭体系中含氧	I1	单键中碘

3. 由原子属性表达的共价键方案

两个不同原子属性的非氢原子通过共价键相互连接。该共价键及其键长可通过这两个原子的原子属性加以确定。确定由两个原子属性表达的化学键与原子属性定义中键的优先级次有关。如

果两原子优先级次不同时，应以优先级次低的为准。如 $C1 - C1$, $C1 - C2$ 或 $C1 - C2$ 都表示碳 - 碳单键 ($C1$ 为准); $C2 - C2$ 与 $C3 - C3$ 则表示形成碳 - 碳双键和碳 - 碳三键; $CC - CC$ 以及 $C2 - CC$ 或 $C3 - CC$ 都表示形成碳 - 碳共轭双键; $CC - OO$ 表示碳 - 氧共轭双键。然而，化学键长的确定是近似的。不仅同一分子中各种化学键的键长不同，而且同一种化学键在不同分子环境中键长也有所变化，加之实验测定所有化学键的键长及其变化几乎是不可能的，因此要准确地估计各种各样化合物分子中的各种化学键的键长是非常困难的，只能采用合理的近似方法。众所周知，除共轭体系与张力大的环体系之外，在大多数情况下，同一种化学键的键长在不同分子环境中的变化不大，可视同一种化学键的键长相同，从而简化计算程序。在共轭体系中，各个键长趋于平均化，即原孤立单键缩短和原孤立双键加长，因此有理由将共轭体系中的单键与双键都近似地看成是单键与双键的平均键长。

应当指出在 MEDV-4 方法中碳 - 碳三键参与共轭时未按共轭处理，而是按三键处理。如在分子碎片“ $C - C \equiv C - C = O$ ”中，原子属性表达为：“ $C1 - C3 \equiv C3 - CC = OO$ ”，则 $C1 - C3$ 按碳 - 碳单键处理（即键长为 $BL = 0.154 \text{ nm}$ ）， $C3 - C3$ 按碳 - 碳三键处理 ($BL = 0.120 \text{ nm}$)， $C3 - CC$ 按碳 - 碳共轭键处理 ($BL = 0.144 \text{ nm}$)， $CC - OO$ 按碳 - 氧共轭键处理 ($BL = 0.137 \text{ nm}$)。

在 MEDV-4 方法理论中应用的各种化学键键长数据取自文献^[4]。常见的由两个不同原子属性的非氢原子确定的化学键及其键长数据参见表 2-3。

表 2-3 有机及药物分子中各种原子属性表达的化学键及其键长

No	化学键	意义	键长/nm
1	$C1 - C1, C1 - C2, C1 - C3$	碳 - 碳单键	0.154
2	$C2 - C2$	碳 - 碳双键	0.134

续表

No	化学键	意义	键长/nm
3	C3 - C3	碳 - 碳三键	0.120
4	CC - CC, C2 - CC, C3 - CC	碳 - 碳共轭双键	0.144 ^a
5	C1 - N1, C2 - N1, CC - N1, C1 - N2	碳 - 氮单键	0.147
6	C2 - N2	碳 - 氮双键	0.130
7	C3 - N3	碳 - 氮三键	0.116
8	CC - NN	碳 - 氮共轭双键	0.142 ^b
9	C1 - O1, C2 - O1	碳 - 氧单键	0.143
10	C2 - O2	碳 - 氧双键	0.122
11	CC - OO	碳 - 氧共轭双键	0.137 ^c
12	NN - OO	氮 - 氧共轭双键	0.122 ^d
13	C1 - S1, C2 - S1	碳 - 硫单键	0.182
14	SS - OO	碳 - 硫共轭双键	0.143 ^e
15	C1 - F1, C2 - F1, C3 - F1, CC - F1	碳 - 氟单键	0.138
16	C1 - L1, C2 - L1, C3 - L1, CC - L1	碳 - 氯单键	0.177
17	C1 - B1, C2 - B1, C3 - B1, CC - B1	碳 - 溴单键	0.193
18	C1 - I1, C2 - I1, C3 - I1, CC - I1	碳 - 碘单键	0.214

a: 按 $\text{C}=\text{C}-\text{C}=\text{C}-\text{C}$ 计算, 取 $(0.154 + 0.134)/2 = 0.144 \text{ nm}$;

b: 按 $\text{C}-\text{C}=\text{N}-\text{C}=\text{C}$ 计算, 取 $(0.134 + 0.147 + 0.130 + 0.154)/4 = 0.141 \text{ nm}$;

c: 按 $\text{C}-\text{C}=\text{C}-\text{C}=\text{O}$ 计算, 取 $(0.154 + 0.134 + 0.154 + 0.122)/4 = 0.141 \text{ nm}$;

d: 按 NO_2 中 $\text{O}=\text{N}=\text{O}$ 计算;

e: 按 SO_3 中 $\text{O}=\text{S}=\text{O}$ 键计算。

4. 相对键长 d 和相对电负性 q

一个化学键的“相对键长 d (relative bond length)”定义为该键键长与碳 - 碳单键键长 (0.154 nm) 的比值。例如, 碳 - 碳单键的

相对键长 $d = 0.154 \text{ nm} / 0.154 \text{ nm} = 1.0000$;而 碳 - 碳双键的相对键长 $d = 0.134 \text{ nm} / 0.154 \text{ nm} = 0.8701$ 。

各个非氢原子的相对电负性定义为该原子 Pauling 电负性与碳原子 Pauling 电负性的比值。常见有机物如药物分子中各非氢元素原子的 Pauling 电负性 (p) 以及相对电负性 (q) 见表 2-4。

表 2-4 几种元素的 Pauling 电负性 (p) 与相对电负性 (q)

元素	p	q	元素	p	q	元素	p	q
C	2.55	1.0000	P	2.19	0.8588	Cl/L	3.16	1.2392
N	3.04	1.1922	S	2.58	1.0118	Br/B	2.96	1.1608
O	3.44	1.3490	F	3.98	1.5608	I	2.66	1.0431

2.1.2 MEDV-4 的构建

定义了原子属性,并规定了由两个不同原子属性的原子表达的共价键及其相对键长,就可计算分子中任意两个通过一个或几个共价键连接的非氢原子(原子 i 和原子 j)之间的相对距离。该距离即为从原子 i 到原子 j 之间的路径中各个化学键的相对键长之和,以 d_{ij} 表示。如果从原子 i 到原子 j 之间的路径不只一条,则规定最短路径中各个化学键之和为相对距离 d_{ij} 。

众所周知,一个分子由各种不同原子属性的原子按不同连接方式(化学键)构成,那么该分子的性质或生物活性就应该是分子中各连接原子之间相互作用的函数。原子之间的相互作用与原子电性及相隔距离密切相关。一般说来,电性越大、距离越小,相互作用就越大。另一方面,不同原子类型之间的相互作用方式也不同。为此,我们提出以基于 4 种原子类型的“分子电性距离矢量(MEDV-4)”来模拟这些非氢原子之间的相互作用,以间接表征其分子结构。

分子电性距离矢量 MEDV-4 中各元素(即分子描述子)值按

如下公式计算^[1]：

$$x_r = m_{kl} = \sum_{i \in k, j \in l} \frac{q_i \cdot q_j}{d_{ij}^2} \quad (k=1,2,3,4; k \leq l \leq 4) \quad (2-1)$$

其中 k 或 l 是各非氢原子的原子类型（共 4 类）原子 i 和 j 分别属于第 k 原子类型和第 l 原子类型； q_i 和 q_j 指原子 i 和 j 的相对电性 d_{ij} 表示原子 i 和 j 之间的相对距离（相对键长表示），是两原子之间各个化学键的相对键长之和，是从原子 i 通过一个或多个化学键连接到原子 j 的所有路径中各个相对键长加和结果的最小值。加和号“ $\sum$ ”表示对所有属于原子类型 k 的非氢原子 i 和属于原子类型 l 的非氢原子 j 进行加和求解。由于 (2-1) 式右端 q_i 和 q_j 为原子 i 和原子 j 的“相对电性”因此我们称由 m_{kl} 表示的矢量描述子为基于 4 种原子类型的“分子电性距离矢量”，简称 MEDV-4 或 MEDV 显然，该矢量中有 10 个元素或称 10 个分子描述子即 $m_{11}, m_{12}, m_{13}, m_{14}, m_{22}, m_{23}, m_{24}, m_{33}, m_{34}, m_{44}$ 或者统一记为： $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$ 。

下面以“2,2-二甲基-3-戊醇”分子为例，说明由 (2-1) 式计算该分子的 MEDV-4 矢量中各描述子变量的详细过程。该分子结构骨架见图 2-2 图中某非氢原子 C_k 或 C_l 表示该编码为 k 的 C 原子属于第 k 原子类型或该编码为 j 的 C 原子属于第 l 原子类型。首先计算各个原子类型原

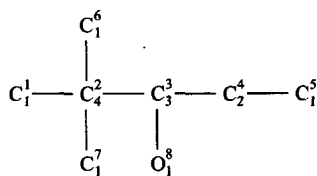


图 2-2 2,2-二甲基-3-戊醇分子骨架图

子之间的距离 d_{ij} 。若某类原子类型不存在于该分子中，则与该原子类型之间的距离视为无穷大，使其相互作用为零或称不发生相互作用。另外，由表 2-4 可知碳与氧的相对电负性分别为 1.000 0 和 1.349 0，碳碳单键与碳氧单键的相对键长分别为 1.000 0 和

0.928 6。那么,有关 x_r, m_{kl}, d_{ij} 等物理量及 MEDV-4 矢量中各描述子的计算结果见表 2-5。

表 2-5 2,2-二甲基-3-戊醇分子中不同 r, k, l 下的距离项 d_{ij} 和 m_{kl}

r	k	j	$d_{ij}(C_k^i - C_l^j)$	m_{kl}
1	1	1	$2(C_1^1 - C_1^6), 2(C_1^1 - C_1^7), 2.928\ 6(C_1^1 - O_1^8),$ $4(C_1^1 - C_1^5), 2.928\ 6(C_1^6 - O_1^8), 2(C_1^6 - C_1^7),$ $4(C_1^6 - C_1^5), 2.928\ 6(C_1^7 - O_1^8), 4(C_1^7 - C_1^5),$ $2.928\ 6(O_1^8 - C_1^5)$	1.5667
2	1	2	$3(C_1^1 - C_2^4), 3(C_1^6 - C_2^4), 3(C_1^7 - C_2^4),$ $1.928\ 6(O_1^8 - C_2^4), 1(C_1^5 - C_2^4)$	1.696 0
3	1	3	$2(C_1^1 - C_3^3), 2(C_1^6 - C_3^3), 2(C_1^7 - C_3^3),$ $0.928\ 6(O_1^8 - C_3^3), 2(C_1^5 - C_3^3)$	2.5645
4	1	4	$1(C_1^1 - C_4^2), 1(C_1^6 - C_4^2), 1(C_1^7 - C_4^2),$ $1.928\ 6(O_1^8 - C_4^2), 3(C_1^5 - C_4^2)$	3.473 8
5	2	2	∞	0
6	2	3	$1(C_2^4 - C_3^3)$	1.000 0
7	2	4	$2(C_2^4 - C_4^2)$	0.250 0
8	3	3	∞	0
9	3	4	$1(C_3^3 - C_4^2)$	1.000 0
10	4	4	∞	0

表 2-5 中各分子描述子的计算如下:

$$x_1 = m_{11} = \frac{1}{2^2} \times 3 + \frac{1}{4^2} \times 3 + \frac{1.349\ 0}{2.928\ 6^2} \times 4 = 1.566\ 7$$

$$x_2 = m_{12} = \frac{1}{1^2} + \frac{1}{3^2} \times 3 + \frac{1.349\ 0}{1.928\ 6^2} = 1.696\ 0$$

$$x_3 = m_{13} = \frac{1}{2^2} \times 4 + \frac{1.349\ 0}{0.928\ 6^2} = 2.564\ 5$$

$$x_4 = m_{14} = \frac{1}{3^2} + \frac{1}{1^2} \times 3 + \frac{1.349\ 0}{1.928\ 6^2} = 3.473\ 8$$

$$x_5 = x_8 = x_{10} = 0.000\ 0$$

$$x_6 = x_9 = \frac{1}{1^2} = 1.000\ 0$$

$$x_7 = m_{24} = \frac{1}{2^2} = 0.250\ 0$$

应当指出,对于组成简单的某些特殊分子,其矢量分子描述子 MEDV-4 可以简化。对于只由碳氢两元素及饱和共价键组成的饱和烃分子体系,因分子中只有一种非氢原子碳,其相对电性等于 1.000 0 即 q_i 与 $q_j = 1$, 另外,所有共价键均为碳碳单键,其相对键长等于 1.000 0 则 d_{ij} 的大小、即是最短路径中所有化学键之加和,此时相对距离简化为拓扑距离。对于烷烃,我们曾提出过几何平均形式的类似矢量,即分子距离边数矢量 λ [2], 并以多元线性回归法和人工神经网络方法成功地应用于烷烃分子的结构表征及其沸点、密度、折光率、热容等物理化学性质及热力学函数之间的定量结构-性质相关 QSPR 研究 [3,5,6]。

对于只含有一个杂原子的有机分子体系, MEDV-4 可简化为分子距边矢量 μ [7]:

$$\mu_r = m_{kl} = \sum_{i \in k, j \in l} \frac{1}{d_{ij}^2} \quad (k=1,2,3,4; k \leq l \leq 4) \quad (2-2)$$

对于不含杂原子的多环芳烃等多种化学键类型的有机分子,在计算 MEDV-4 时不必计算相对电性(因为所有非氢原子的相对电负性均等于 1),只需计算相对距离即可。此时 MEDV-4 变换为(2-3)式 [8]:

$$x_r = m_{kl} = \sum_{i \in k, j \in l} \frac{1}{d_{ij}^2} \quad (k=1,2,3,4; k \leq l \leq 4) \quad (2-3)$$

2.1.3 MEDV-4 的计算程序

对于较为复杂的有机或药物分子，其 MEDV-4 的计算量还是比较大的，为了减轻手工计算工作量，特别是消除手工计数距离时容易造成的差错（如重复计数或少计数），应该也必须采用计算机技术。作者开发了基于 True BASIC 语言^[9,10]的计算程序“GITMEDV”，用于计算分子电性距离矢量 MEDV-4。程序 GITMEDV 分为结构输入、邻接矩阵与距离矩阵求解、MEDV 计算及数据输出四个部分。其程序流程简图如图 2-3 所示。

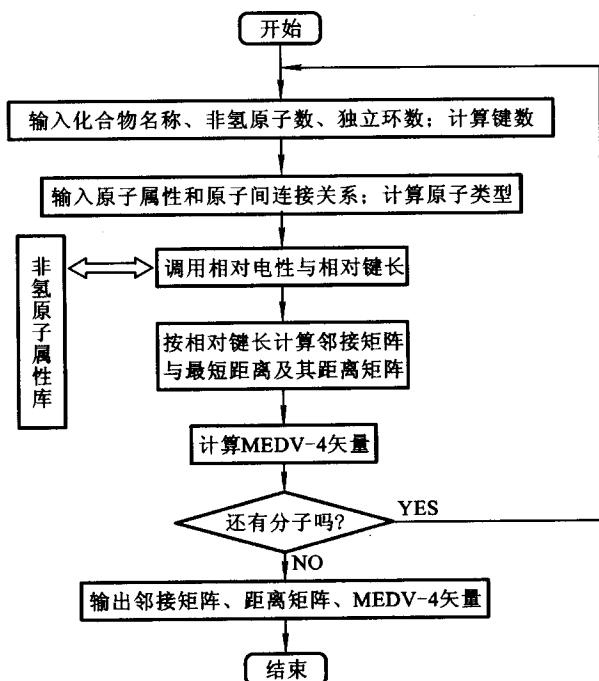


图 2-3 分子电性距离矢量 (MEDV-4) 计算流程图

在“结构输入”部分,首先给定待计算分子的初始化数据,包括分子的名称、分子所含非氢原子的数目 (n_{H}) 以及分子中存在的独立环的个数 (n_{R})。然后,以字符串的形式输入分子中各个非氢原子的原子属性;接着按分子中各非氢原子编码顺序,输入分子中各原子属性表达的所有化学键两端的非氢原子编码序号,即非氢原子之间的相互连接关系,以数字形式输入。这是待计算 MEDV-4 分子描述子的“骨架分子”与用户直接对话的部分,是对分子二维拓扑结构以及原子特性进行数值化的最原始步骤。

例如 图 2-2 所示“2,2-二甲基-3-戊醇”分子在程序 GITMEDV 中的输入部分的 True BASIC 语言的源代码为:

DATA	“2,2-二甲基-3-戊醇”,8,0	! 名称,非氢原子数,环个数
DATA	“C1”,“C1”,“C1”,“C1”,“C1”, “C1”,“C1”,“O1”	! 各原子属性
DATA	1,2,2,3,2,6,2,7,3,4,3,8,4,5	! 7 个化学键两端的原子序号

在这个源代码中,初始化数据为分子名称“2,2-二甲基-3-戊醇”非氢原子个数为 8 :即“ $n_{\text{H}}=8$ ”独立环个数为零 即“ $n_{\text{R}}=0$ ”(源代码的第 1 行)。该分子的结构输入部分包括:以原子编码顺序输入的 8 个非氢原子的原子属性,即:“C1”,“C1”,“C1”,“C1”,“C1”,“C1”,“C1”,“O1”(源代码的第 2 行)以及这 8 个原子的相互连接关系,即 7 个化学键连接的非氢原子的编码序号序列:1,2,2,3,2,6,2,7,3,4,3,8,4,5(源代码的第 3 行)。这种输入的一个好处是:该序列中相同原子编码序号的个数就是该编码序号对应原子的原子类型。比如该序列中有四个“2”表示编号序号为 2 的碳原子属于第 4 原子类型。而 1,5,6,7,8 号原子则都属于第 1 原子类型。

“邻接矩阵与距离矩阵计算”部分的计算结果不仅可以为下一

步计算 MEDV-4 分子描述子作准备, 而且也可以为进行有关计算与比较基于距离与邻接矩阵的大量拓扑指数提供原始计算数据。应当指出, 在 GITMEDV 中, 邻接矩阵和距离矩阵都是按相对键长标准计算相对距离进行设计的, 而不是按一般拓扑指数中定义的拓扑距离 (一个化学键的长度为 1) 为标准设计的。然而, 只要在程序中指定所有化学键的距离都等于 1 即可转化为一般意义下的邻接矩阵和距离矩阵。

对于含有一个或多个环的分子体系, 从原子 i 到原子 j 的路径一般有多条, 那么其距离的计数也有多个。在计算 MEDV-4 的 GITMEDV 程序中, 选择最短路径作为两个原子之间的距离量度。也即 MEDV-4 方法中的距离矩阵是以最短相对距离为标准的。两个非氢原子之间最短距离的计算方法采用文献[11]中的最短距离算法^[11]。

“MEDV-4 计算”部分从上一步得到的最短距离表达的距离矩阵出发, 结合原子类型并利用各种非氢原子的相对电负性, 代入 (2-1) 式计算 MEDV-4 矢量中的 10 个分子描述子。

GITMEDV“输出”部分可以按需要选择输出邻接矩阵、距离矩阵、MEDV 矢量等数据格式数据, 供有关文字处理软件直接调用。也可以按数据子程序库的形式输出, 以直接供其他程序如定量结构-活性相关 (QSAR) 分析程序调用。

2.2 分子全息距离矢量 (MHDV)

“分子全息距离矢量”^[12,13] 是 MEDV-4 的改进版, MHDV 是英文 “molecular holographic distance vector” 的简称。MEDV-4 是基于 4 种原子类型的, 不管该原子是碳原子, 还是氮原子等其他非氢原子, 都没有加以区分。这个新的矢量型分子描述子 MHDV 在保留原 MEDV-4 中定义的原子属性的基础上, 进一步区分 N, O, F 等杂原子, 而将非氢原子划分为 13 种原子类型, 以期提高对分子结构的分

分辨率。另一方面 原 MEDV-4 中定义的原子属性包括了原子本身的电负性和化学键特性 而在 MHDV 中则只含有化学键的特性, 其原子电负性属性则归入原子类型的划分方案中, 此时所有非氢原子的相对电负性均规定为 1。也就是说 MHDV 不考虑非氢原子的元素电负性, 但更充分地考虑各种原子属性表达的化学键即各种距离。这也是命名“全息距离”的原因。通过对结构复杂且杂原子较多的两组二肽分子的 QSAR 研究表明, 对于原子多样性的二肽分子体系其结果确实大大优于 MEDV-4 方法。

2.2.1 原子类型、原子属性、相对键长

不同于 MEDV-4 方法, 在 MHDV 方法中, 原子类型 (atomic type) 不仅与所连非氢原子有关, 而且还考虑非氢原子本身的元素特性。这个“原子类型”定义为与该原子相连的非氢原子个数加上该原子的识别特性或识别号 (ID)。各非氢原子的识别号定义为: $ID = 4 \times (v - 4)$ 式中 v 为价电子层中电子数。由 ID 定义可知, 非氢原子 C, N, P, O, S, F, Cl, Br, I 等原子的识别号 ID 值分别为 0, 4, 4, 8, 8, 12, 12, 12, 12。例如, 色氨酸分子 (分子结构参见图 2-4) 中 15 个编码非氢原子 (从小到大顺序) 的原子类型分别为: 5, 3, 3, 9, 9, 2, 3, 3, 2, 2, 2, 2, 3, 6, 2。

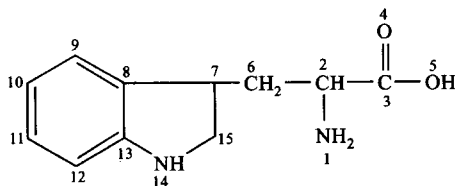


图 2-4 色氨酸分子结构中
各非氢原子的编码序号

MHDV 方法 13 种原子类型的划分方案如表 2-6。

表 2-6 MHDV 中的 13 原子类型分类方案

类型	原子环境*	ID	类型	原子环境*	ID	类型	原子环境*	ID
1	C—	0	6	—N—, —P—	4	11	>S—	8
2	—C—	0	7	>N—, >P—	4	12	>S<	8
3	>C—	0	8	>N<, >P<	4	13	F—, Cl—, Br—, I—	12
4	>C<	0	9	O—, S—	8			
5	N—, P—	4	10	—O—, —S—	8			

* —, >, < 分别表示连有 1, 2, 2 个非氢原子或化学键。

MHDV 中“原子属性”仍采用 MEDV-4 中的双字符串表示, 两字符的意义也与前 MEDV-4 中的原子属性相同, 但不考虑电负性变化。即 OO, C2 分别表示与其他非氢原子形成一个共轭双键的氧原子和形成一个双键的碳原子。

MHDV 方法中“相对键长”仍定义为该键键长与碳碳单键键长的比值。

2.2.2 MHDV 的构建

按照新的 13 种原子类型划分方案, 仅考虑分子中各原子之间的全部距离, 而不考虑各原子电负性的变化, 即令所有非氢原子电负性为 1 则 MHDV 矢量中分子描述子的计算公式定义如下:

$$x_r = h_{kl} = \sum_{i \in k, j \in l} \frac{1}{d_{ij}^2} \quad (k=1, 2, 3, \dots, 13; k \leq l \leq 13) \quad (2-4)$$

其中 k 或 l 同样是各非氢原子的原子类型, 但 MHDV 中有 13 种原子类型, 而不是原 MEDV 中的 4 种原子类型; 原子 i 和 j 分别属于第 k 原子类型和第 l 原子类型; d_{ij} 表示原子 i 和 j 之间的相对距离, 是从原子 i 到原子 j 的各个路径中化学键相对键长加和的最小值。加和号“ Σ ”表示对所有属于原子类型 k 的非氢原子 i 和属于原子类型 l 的非氢原子 j 求和。由于 (2-4) 式右端已无“相

对电性”而只有基于 13 种原子类型的更加全面的距离 (d_{ij}) 信息 因而称 h_{kl} 为“分子全息距离矢量”简称 MHDV。由于 MHDV 方法定义了 13 种原子类型, 因而 MHDV 矢量中共有 $(13 + 1) \times 13/2 = 91$ 个描述子。在 MHDV 中, r, k, l 三者的对应关系参见表 2-7。

表 2-7 MHDV 计算公式中 $r, k, l (x_r = h_{kl})$ 三者的相互关系

r	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$	$l=6$	$l=7$	$l=8$	$l=9$	$l=10$	$l=11$	$l=12$	$l=13$
$k=1$	1	2	3	4	5	6	7	8	9	10	11	12	13
$k=2$		14	15	16	17	18	19	20	21	22	23	24	25
$k=3$			26	27	28	29	30	31	32	33	34	35	36
$k=4$				37	38	39	40	41	42	43	44	45	46
$k=5$					47	48	49	50	51	52	53	54	55
$k=6$						56	57	58	59	60	61	62	63
$k=7$							64	65	66	67	68	69	70
$k=8$								71	72	73	74	75	76
$k=9$									77	78	79	80	81
$k=10$										82	83	84	85
$k=11$											86	87	88
$k=12$												89	90
$k=13$													91

对于含杂原子种类与数目较多, 化学键类型比较复杂的分子体系 计算 MHDV 中各个描述子变量是相当费时、相当麻烦和容易出错的, 只有也必须采用计算机技术才能多快好省, 提高效率。对原计算程序 GITMEDV 稍做修改就可以完成 MHDV 方法所规定的各项计算工作。这个改进后的 True BASIC 语言计算程序更

名为“GITMHDV”。具体计算实例参见第五章。

2.3 基于 13 种原子类型的 MEDV

基于 13 种原子类型的 MEDV 简称 MEDV-13 也是 MEDV-4 的改进版^[14-17]。比 MHDV 方法更进一步, 在应用 MHDV 中 13 种原子类型的同时, 还考虑各原子相对电负性变化。在该方法中, 应用修正的电拓扑状态指数 (简称 E-状态指数) 替换 MEDV-4 中的相对电性, 同时以拓扑距离代替 MEDV-4 中由相对键长表达的相对距离。引入拓扑距离是为了简化距离的计算程序, 而引入修正的 E-状态指数是为了表征同一原子在不同分子环境中相对电性会发生变化之事实。

2.3.1 原子类型与原子属性

MEDV-13 中“原子类型”采用 MHDV 方法中定义的 13 种类型。即原子类型等于与之连接的非氢原子数加上该原子本身的识别号 ID (参见表 2-6)。

MEDV-13 方法以“拓扑距离”即化学键数代替 MEDV-4 和 MHDV 方法中的相对距离, 因而距离计数更加简单。但在两原子之间存在多个路径时, 仍采用最短距离标准。由于不用相对距离, 故 MEDV-4 与 MHDV 中的“原子属性”概念在 MEDV-13 中不再使用。

MEDV-13 方法中新定义的“原子属性”概念类似于 E-状态指数方法中的原子类型概念。E-状态指数由 Hall 和 Kier 等提出并发展, 已广泛地用于有机及药物分子理化性质与生物活性的 QSPR/QSAR 分析^[18,19]。作者参照 Hall 和 Kier 等提出的方法, 对含有 C, N, O, S, P, F, Cl, Br, I 等非氢原子的药物分子所属原子属性进行划分, 共计 43 类。具体划分结果及有关参数均列于表 2-8 中。

表 2-8 电拓扑状态, 43 个原子属性, 13 种原子类型和原子固有状态 I

属性	原子	类型	N	σ	π	n	h	v	I
1	$-\text{CH}_3$	1	2	4	0	0	3	4	2
2	$-\text{CH}_2-$	2	2	4	0	0	2	4	1.5
3	$-\text{CH}<$	3	2	4	0	0	1	4	1.333 3
4	$>\text{C}<$	4	2	4	0	0	0	4	1.25
5	$=\text{CH}_2$	1	2	3	1	0	2	4	3
6	$=\text{CH}-$	2	2	3	1	0	1	4	2
7	$=\text{C}<$	3	2	3	1	0	0	4	1.666 7
8	$=\text{C}=\text{}$	2	2	2	2	0	0	4	2.5
9	$\equiv\text{CH}$	1	2	2	2	0	1	4	4
10	$\equiv\text{C}-$	2	2	2	2	0	0	4	2.5
11	cCH_2	1	2	3	0.5	0	2	4	2.5
12	$\text{cCH}-$	2	2	3	0.5	0	1	4	1.75
13	$\text{cC}<$	3	2	3	0.5	0	0	4	1.5
14	cCHc	2	2	3	1	0	1	4	2
15	$-\text{cCc}$	3	2	3	1	0	0	4	1.666 7
16	cCcc	3	2	3	1.5	0	0	4	1.833 3
17	$-\text{OH}$	9	2	2	0	4	1	6	2.449 5
18	$-\text{O}-$	10	2	2	0	4	0	6	1.837 1
19	$=\text{O}$	9	2	1	1	4	0	6	3.674 2
20	cO	9	2	1	0.5	4	0	6	3.061 9
21	$-\text{NH}_2$	5	2	3	0	2	2	5	2.236 1
22	$-\text{NH}-$	6	2	3	0	2	1	5	1.677 1
23	$>\text{N}-$	7	2	3	0	2	0	5	1.490 7
24	$=\text{NH}$	5	2	2	1	2	1	5	3.354 1
25	$=\text{N}-$	6	2	2	1	2	0	5	2.236 1
26	$\equiv\text{N}$	5	2	1	2	2	0	5	4.472 1

续表

属性	原子	类型	N	σ	π	n	h	v	I
27	cNH	5	2	2	0.5	2	1	5	2.795 1
28	cN—	6	2	2	0.5	2	0	5	1.956 6
29	cNc	6	2	2	1	2	0	5	2.236 1
30	$\geq N =$	7	2	3	2	0	0	5	2.236 1
31	—SH	9	3	2	0	4	1	6	1.769 1
32	—S—	10	3	2	0	4	0	6	1.156 7
33	$=S$	9	3	1	1	4	0	6	2.313 4
34	$>S=$	11	3	3	1	2	0	6	1.13 4
35	$\geq S \leq$	12	3	4	2	0	0	6	1.122 7
36	—F	13	2	1	0	6	0	7	2.645 8
37	—Cl	13	3	1	0	6	0	7	1.910 8
38	—Br	13	4	1	0	6	0	7	1.653 6
39	—I	13	5	1	0	6	0	7	1.534 5
40	—PH ₂	5	3	3	0	2	2	5	1.614 9
41	—PH—	6	3	3	0	2	1	5	1.055 9
42	$>P—$	7	3	3	0	2	0	5	0.869 6
43	$\geq P <$	8	3	4	1	0	0	5	0.900 6

表 2-8 中符号“— = $\equiv$, $>$, $<$, c, cc, $\geq$, $\leq$ ”分别表示一个单键、一个双键、一个三键、两个单键、两个单键、一个共轭双键、两个共轭双键、一个双键和一个单键、一个双键和一个单键等。例如, 表中原子属性 35 即“ $\geq S \leq$ ”表示该硫原子与 4 种非氢原子形成 2 个双键和 2 个单键。

与文献^[19]不同的是, 我们的方法对芳香环没有划分新的属性。考虑到共轭体系中 π 电子为多个原子所共有, 这与一般双键中 π 电子是不一样的, 因此对处于共轭体系两端或中央的各非氢原子的原子属性加以区分。自然有关芳香环体系就进入了共轭体

系的属性范围之中。表 2-8 中原子属性 27 与 28 即“cNH”和“cN-”)分别表示处于共轭体系端头的氮原子；而原子属性 29 即“cNc”)则表示共轭体系中间的氮原子。

2.3.2 相对电性量度：修正的 E-状态指数

元素原子的 Pauling 电负性是不随原子在不同分子环境的变化而变化的，即不论在什么样的分子环境中，某元素原子的电负性是不变的。这不能反映同一原子之电性在不同分子化学环境中的变化情况。因此，我们引入 Hall 和 Kier 的 E-状态指数^[24]并稍作修改，以表征不同分子中原子电负性的变化。

首先引入原子固有状态 (I) 概念。我们将 I 定义为：

$$I = \sqrt{v/4} [(2/n)^2 \delta^v + 1] / \delta \quad (2-5)$$

式中 v 是原子价电子层的电子数； n 是该原子价电子层所属主量子数； δ^v 和 δ 是原子在不同分子环境中构成不同共价键时的分子连接性 δ 值：

$$\delta = \sigma - h \quad \delta^v = \sigma + \pi - h \quad (2-6)$$

式中 σ 和 π 分别是该原子参与 σ 轨道和 π 轨道成键的电子数； h 是与该原子相连的氢原子个数。上述定义与原 Hall 和 Kier 在 E-状态指数方法中的定义相比，(2-5)式多了一个系数 $(v/4)^{1/2}$ ，而(2-6)式中 δ^v 的计算则没有包含孤对电子。

由于孤对电子没有参与共价键的形成，因而未计入 δ^v 中。也正因为如此，同一周期中与非氢原子形成相同 σ 键和 π 键的各元素原子属性（比如—CH₃，—NH₂，—OH 和—F 等）将具有相同的 δ^v 和 δ 值。为使处于同一周期不同价电子状态的原子具有不同的固有状态 I 值，我们在原子固有状态 I 中引入了系数 $(v/4)^{1/2}$ 。由于有机分子多以碳原子为分子骨架，系数 4 即是碳原子的价电子数。由上可知，原子固有状态 I 是与原子在分子中的具体化学

环境即参与形成的化学键类型及数目密切相关的。定义原子属性概念的基本规则是：相同原子属性具有相同 I 值，不同原子属性具有不同 I 值。

在原子属性的分类方案中，共轭体系中 π 电子为多个原子所共有，这与一般双键中的 π 电子是不一样的。因此对于共轭体系，在计数 δ^v 中 π 电子时应作减半处理。例如，表 2-8 中原子属性 12 即“ cCH- ”的 δ 与 δ^v 计算如下：

$$\delta = \sigma - h = 3 - 1 = 2; \quad \delta^v = \sigma + \pi - h = 3 + 0.5 - 1 = 2.5$$

有了修正的原子属性划分方案及相应的原子固有状态 (I)，作为相对电性 (q_i) 量度的电拓扑状态指数即 E-状态指数 (q_i) 定义为：

$$q_i = I_i + \sum_{j \neq i}^{\text{all } j} (I_i - I_j) / d_{ij}^2 \quad (2-7)$$

式中 d_{ij} 为两原子 i 与 j 之间的最短拓扑距离。即从原子 i 到原子 j 的所有路径中化学键数加和的最小值。

2.3.3 MEDV-13 的构建

以修正的 E-状态指数 q_i 代替相对电负性，拓扑距离 d_{ij} 代替相对距离 应用 13 种原子类型，则 MEDV-13 的计算公式可写为类似 MEDV-4 的计算公式 (2-1) 的形式：

$$x_r = m_{kl} = \sum_{i \in k, j \in l} \frac{q_i q_j}{d_{ij}^2} \quad (k, l = 1, 2, 3, \dots, 13; l \geq k; r = 1, 2, 3, \dots, 91) \quad (2-8)$$

式中 k 或 l 是各非氢原子的原子类型，但 MEDV-13 方法中已将原 MEDV-4 中的 4 原子类型扩展为 13 原子类型 原子 i 和 j 分别属于第 k 原子类型和第 l 原子类型； d_{ij} 不像 MEDV-4 与

MHDV 表示最小相对距离，而是表示原子 i 和 j 之间的最短拓扑距离 即从原子 i 到原子 j 的各个路径中化学键数加和的最小值或者说最少化学键数。加和号 “ Σ ” 表示对所有属于原子类型 k 的非氢原子 i 和属于原子类型 l 的非氢原子 j 求和； q_i 与 q_j 是原子在实际分子环境中的修正 E - 状态指数。虽然它也是一种电负性的反映，但明显不同于 Pauling 电负性，它是随原子在不同分子环境中的变化而变化的。

由于引入了 E - 状态指数 计算 MEDV - 13 需要对原 MEDV - 4 计算程序作较大修改。作者在 Visual BASIC 语言平台上，在原 True BASIC 编制的计算程序基础上，开发设计了面向对象的计算 MEDV - 13 各个描述子的应用程序模块，并将其接入我们开发的 VSMP 应用软件中（见第三章）。该程序模块大致分为三个部分：结构输入文件编写、MEDV - 13 分子描述子计算和结果输出等。该模块的基本程序流程如图 2 - 5 所示。

“结构输入文件编写”部分首先给定待计算分子的初始化数据，即分子名称、分子所含非氢原子数（ n_{NH} ）以及分子中存在的独立环个数（ n_{R} ）；然后，针对待编码分子结构中的各非氢原子进行编号，并根据表 2 - 8 中的原子属性划分方案，输入各编号原子的原子属性值；最后按从小到大的顺序输入各非氢原子的连接关系，即所有化学键两端的非氢原子编码序号。与 MEDV - 4 与 MHDV 方法中原子属性的字符型输入不同，MEDV - 13 方法中原子属性的输入是数字型的，即 1~43 的原子属性值。

下面以图 2 - 6 所示三个联苯分子为例，说明计算 MEDV - 13 描述子的结构输入文件。

首先，写入要计算多少个分子的 MEDV - 13 然后依次编写各个分子的分子名称、非氢原子数、独立环数，各非氢原子的原子属性值，分子中各原子的连接关系等。

那么，完整的结构输入文件如下所示：

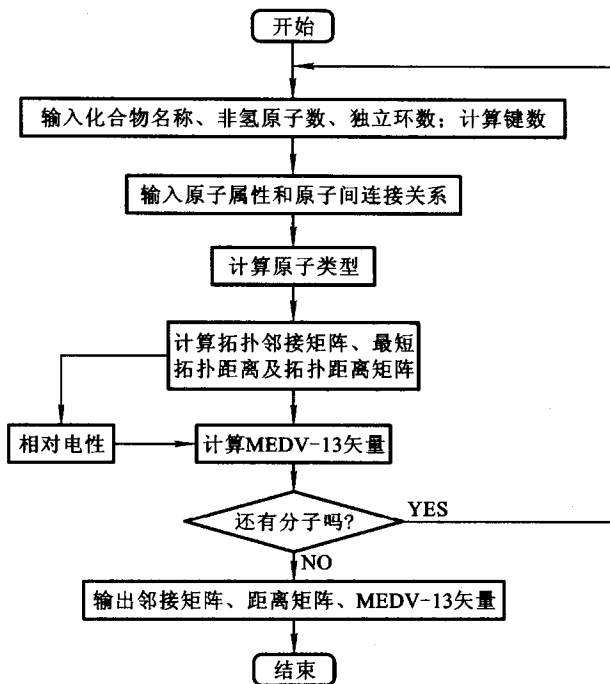


图 2-5 基于 13 原子类型的分子电性距离
矢量 MEDV-13 计算流程图

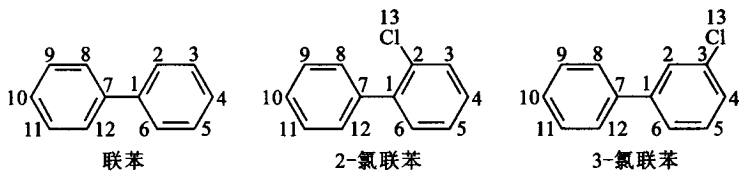


图 2-6 联苯、2-氯联苯及 3-氯联苯的分子骨架结构及原子编号

3	! 要计算的分子数
" Biphenyl",12,2	! 分子名称,非氢原子数,独立环数
16,14,14,14,14,14,16,14,14,14,14,14	! 12 个非氢原子的原子属性,逗号分隔
1,2,1,6,1,7,2,3,3,4,4,5,5,6,7,8,7,12,8,9,9,10,10,11,11,12	! 12 个非氢原子的连接关系,逗号分隔
" 2 - Cl",13,2	! 第二个分子
16,15,14,14,14,14,16,14,14,14,14,14,37	
1,2,1,6,1,7,2,3,2,13,3,4,4,5,5,6,7,8,7,12,8,9,9,10,10,11,11,12	
" 3 - Cl",13,2	! 第三个分子
16,14,15,14,14,14,16,14,14,14,14,14,37	
1,2,1,6,1,7,2,3,3,4,3,13,4,5,5,6,7,8,7,12,8,9,9,10,10,11,11,12	

在上述编码文件中,分子名称是以字符串型的形式输入的,要带双引号。非氢原子连接关系是按各分子结构中的编号顺序从小到大输入的。比如对第 1 个分子 先输入第 1 号原子(骨架结构中的编号)的连接关系(1,2,1,6,1,7) 然后输入第 2 号原子的连接关系(2,3) 而不是第 6 号或其他原子的连接关系。因为只有这样才能不会丢失或重复输入连接关系。同时也不要将连接关系(2,3) 输入成(3,2)。

“MEDV-13 描述子计算”部分应用编码文件输入的数据首先计算各个分子的拓扑邻接矩阵与拓扑距离矩阵,然后利用拓扑距离矩阵和各原子的相对电性即修正的 E -状态指数,并结合 13 种原子类型,根据 2-8) 式计算各个 MEDV-13 分子描述子。

“结果输出”部分以文本文件格式输出各个分子的 MEDV-

13 描述子值。

参 考 文 献

- [1] 刘树深, 刘堰, 李志良, 蔡绍哲. 一个新的分子电性距离矢量 (MEDV). 化学学报, 2000, 58(11): 1353~1357
- [2] Liu S S, Cao C Z, Li Z L. Approach to estimation and prediction for normal boiling point (NBP) of alkanes based on a novel molecular distance-edge (MDE) vector. J. Chem. Inf. Comput. Sci., 1998, 38(3): 387~394
- [3] Liu S S, Cai S X, Cao C Z, Li Z L. Molecular electronegative distance vector (MEDV) relating to 15 properties of alkanes. J. Chem. Inf. Comput. Sci., 2000, 40(6): 1337~1348
- [4] Aylward G H, Findlay T J V. SI Chemical Data. 2nd Edn. 周宁怀译. SI 化学数据表. 北京: 高等教育出版社, 1985; New York: John Wiley & Sons, 1982
- [5] 刘树深, 曹晨忠, 杨万平, 李志良. 改良人工神经网络方法预测烷烃物理性质. 化工学报, 1998, 49(2): 245~250
- [6] Liu S S, Cao C Z, Li Z L. A novel molecular distance-edge (MDE) vector & thermodynamical properties of alkanes. Chem. Hong Kong, 1998, 1(2): 113~123
- [7] Liu S S, Liu H L, Xia Z N, Cao C Z, Li Z L. Molecular distance-edge vector (μ): an extension from alkanes to alcohols. J. Chem. Inf. Comput. Sci., 1999, 39(6): 951~957
- [8] Liu S S, Yin C S, Cai S X, Li Z L. Structural vector description and estimation of normal boiling points for 66 aromatic hydrocarbons. Chin. Chem. Lett., 2000, 11(12): 1089~1092
- [9] 谭浩强, 张基温. True BASIC 程序设计. 第2版. 北京: 清华大学出版社, 1994
- [10] 李世民, 潘晓今, 杨卫东, 刘华红. True BASIC 程序设计教程. 北京: 北京工业大学出版社, 1990
- [11] 许卓群, 张乃孝, 杨冬青, 唐世渭. 数据结构. 北京: 高等教育出版社,

- [12] Liu S S, Yin C S, Cai S X, Li Z L. A novel MHDV descriptor for dipeptide QSAR studies. *J. Chin. Chem. Soc.*, 2001, 48(2): 253 ~ 260
- [13] Liu S S, Yin C S, Wang X D, Wang L S. QSAR studies on dipeptides based on a combinatorial MHDV - GA - MLR method. *J. Chin. Chem. Soc.*, 2002, 49(6): 1089 ~ 1096
- [14] Liu S S, Yin C S, Cai S X, Li Z L. QSAR study of steroid benchmark and dipeptides based on MEDV - 13. *J. Chem. Inf. Comput. Sci.*, 2001, 41(2): 321 ~ 329
- [15] Liu S S, Yin C S, Shi Y Y, Cai S X, Li Z L. MEDV - 13 for QSAR studies on the COX - 2 inhibition by indomethacin amides and esters. *Chin. J. Chem.*, 2001, 19(8): 751 ~ 756
- [16] Liu S S, Cui S H, Shi Y Y, Wang L S. Novel variable selection and modeling method based on the prediction for QSAR of cyclooxygenase - 2 inhibition by thiazolone and oxazolone series. *Internet Electron. J. Mol. Des.*, 2002, 1(11): 610 ~ 619
- [17] Liu S S, Liu H L, Shi Y Y, Wang L S. QSAR of cyclooxygenase - 2 (COX - 2) inhibition by 2, 3 - diarylcyclopentenones based on MEDV - 13. *Internet Electron. J. Mol. Des.*, 2002, 1(6): 310 ~ 318
- [18] Hall L H, Kier L B. Electrotological state indexes for atom types: a novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.*, 1995, 35: 1039 ~ 1045
- [19] Hall L H, Kier L B, Brown B B. Molecular similarity based on novel atom type electrotological state indices. *J. Chem. Inf. Comput. Sci.*, 1995, 35: 1074 ~ 1080

第3章

矢量描述子研究方法

著名拓扑化学家 Randic 在研究各种拓扑指数^[1~5]时曾经指出，一个新的拓扑指数必须至少具有对各种异构体的良好分辨率和与物理化学性质的良好相关性。同样，一个矢量描述子也至少应具有上述性能，才能得以广泛应用。如果认为拓扑指数等一系列单参数分子描述子可作为一维空间来处理的话，那么本文提出的 MEDV-4、MHDV 及 MEDV-13 等矢量结构描述子则必须在多维空间中加以描述。如何研究这些矢量描述子对各种异构体的分辨率以及它们与有关物质分子理化性质 / 生物活性之间的相关性，正是这一章要解决的主要问题。

在这一章，首先建立矢量描述子对分子结构分辨率的方法；然后讨论分子电性距离矢量等矢量描述子在定量构效关系研究中的变量选择问题，建立基于交互检验为目标函数的变量选择方法；进而给出基于线性回归、主成分回归与偏最小二乘回归方法的变量选择与模型化方法；最后讨论模型质量评价问题并简单介绍我们实验室开发的应用软件 VSMP。

3.1 矢量描述子对分子结构的分辨率

分子描述子特别是矢量分子描述子对各个分子特别是同分异构体分子的分辨能力是表征该分子描述子性能的重要指标之一。如果没有对各种异构体良好的分辨能力,也就失去了用于表征分子结构,进而进行分子性能或生物活性估计与预测以及新分子设计的优势。这样的描述子也就失去了研究与应用的意义。

对于单参数分子描述子或者标量描述子,比如许多拓扑指数^[6-10]等只要观其数值大小与差异即可评价其分辨能力。然而,对于矢量分子描述子,仅仅比较其中某个描述子变量的大小与差异是不够的,必须对整个矢量进行整体合理评价。

本章借用模式识别中模式空间的概念^[11, 12]将每一个样本分子的描述子矢量(该矢量有 r 个分子描述子)看成是由 r 个描述子变量张成的 r 维模式空间中的一个点,则两个样本分子描述子矢量点在该模式空间中的距离(简称模式距离 D)可作为这两个分子结构相似性的量度。运用所有各对描述子矢量之间的相似性,可以整体说明矢量描述子对各种异构体结构的分辨情况。两个样本分子描述子矢量点之间的模式距离越大,相异性越大,而相似性则越小,两者越容易分辨。相反,模式距离越小,相异性越小,相似性则越大,两者越难分辨。具有相同分子描述子矢量的两个样本分子之间的模式距离等于零。

定义两个样本分子 i 和 j 在 r 个矢量描述子张成的 r 维模式空间中的模式距离 D_{ij} 为:

$$D_{ij} = [(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ir} - x_{jr})^2]^{1/2} \quad (3-1)$$

式中 $x_{ik}, x_{jk} (k=1, 2, \cdots, r)$ 就是第二章中定义的第 i 分子和第 j 分子的 MEDV 矢量中的各个描述子。对于 MEDV-4 矢量,最多有 10 个分子描述子 ($r=10$); 对于 MHDV 和 MEDV-13 矢量,最

多有 91 个分子描述子 ($r = 91$)。

要整体说明 MEDV-4 等各描述子矢量对数据集中各样本分子结构的分辨能力, 应该计算这组分子中所有样本分子描述子矢量两两之间的模式距离。只有当所有描述子矢量对之间的模式距离都不等于零 (或都大于某个给定的最小正值) 时, 才能说明该分子描述子矢量对这一组分子具有完全的分辨率。此时, 我们说分辨率等于 1 如果所有距离都等于零, 则分辨率等于零。据此定义分辨率 R_D 为:

$$R_D = 1 - \frac{N_E}{N} \quad (3-2)$$

式中 N_E 是指模式距离等于零或小于等于某个给定正数 E 的独立矢量对数; N 是数据集中样本分子对数。那么, 由 (3-2) 式定义的分辨率取值将分布在 0~1 之间。 R_D 越接近于 1 分辨率就越好, R_D 越接近于 0 分辨率就越差。

对于数据集中样本数 (N 较大的分子体系 比如 150 个分子, 其总距离数或描述子矢量对数将是 11175 ($150 \times 149/2 = 11175$), 这个数值是相当大的。为了直观地表述这些距离的分布, 观察有关描述子矢量对这些分子的整体分辨情况, 可将这些距离按大小范围划分为几组, 计数各个组中的距离个数占总距离数的百分率 (频率), 作出频率随距离分布的直方图。例如, 作者在研究 MEDV-4 矢量对 150 个烷烃分子结构的分辨率时, 得到如表 3-1 所示对应 MEDV-4 矢量的模式距离与相应频率分布, 由此所作直方图如图 3-1 所示^[13]。若给定 $E = 0.05$ 则 MEDV-4 对这 150 个烷烃分子结构具有完全的分辨率 ($R_D = 1$)。有关 MEDV-4、MHDV 及 MEDV-13 对各种类型化合物分子结构的分辨情况将在后续有关矢量分子描述子的应用章节中具体讨论。

表 3-1 150 个烷烃中 MEDV-4 的模式距离与频率

模式距离区间	频率/%
0.00~0.05	0.00
0.05~0.50	0.40
0.50~1.50	2.17
1.50~2.50	5.73
2.50~3.50	9.70
3.50~4.50	17.58
4.50~5.50	20.47
5.50~6.50	17.05
6.50~7.50	10.20
7.50~8.50	6.95
8.50~9.50	5.59
9.50~10.50	2.68
10.50~11.50	0.91
11.50~12.50	0.49
12.50~13.50	0.04
13.50~14.50	0.02

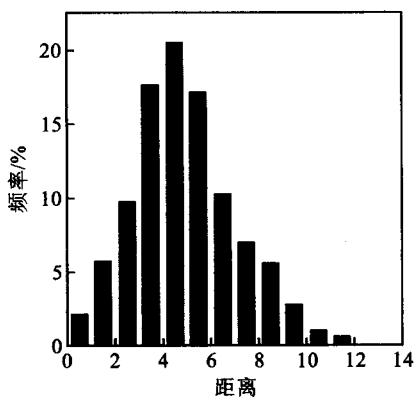


图 3-1 150 个烷烃 MEDV-4 模式距离频率分布

上述 MEDV - 4 矢量对 150 个烷烃的分辨率结果表明, MEDV - 4 能够区分烷烃分子中的各种结构异构体, 优于目前绝大多数单一的拓扑指数的分辨率。在这 150 个烷烃中包括了所有 1~10 个碳原子构成的全部异构体, 这说明 MEDV - 4 能反映烷烃中的直链烷, 也能表征各种支链烷烃的结构, 即包含了分子大小的结构, 也包含有关长度与体积的结构信息, 因而与烷烃的物理化学性质具有良好的相关性, 具体应用实例参见第四章 MEDV 在 QSPR 研究中的应用内容。尽管 MEDV 只是基于分子拓扑结构与原子电性性质(在烷烃分子中只有拓扑距离信息), 但与许多基于分子距离矩阵或邻接矩阵的拓扑指数不同, MEDV 是一个矢量而不是标量, 因而保留了原有分子更多的结构信息, 具有比单一指数高得多的分辨率。这也正是矢量描述子的优势之一。

3.2 变 量 选 择

对于某一具体的化合物体系, 矢量描述子中的某些变量可能是重要的, 而另一些描述子则可能是无关紧要的或者说对所研究的某些性质的贡献不大。因此, 在应用矢量描述子进行定量分子结构 - 物理化学性质 / 生物活性相关 QSPR/QSAR 研究时必须考虑矢量描述子中的变量选择问题。另一方面, 即使不是矢量描述子, 而是某些不同来源的分子结构参数(比如从各种结构参数库中选择的各种分子结构描述子或其他参数)的集合, 也会因其中某些参数统计意义不大, 或者变量之间存在共线性关系等, 如不消除这些变量的影响就可能会大大降低模型统计的显著性, 因此, 也必须进行合理的变量选择与优化。

尽管作者提出的 MEDV - 4 等三种分子电性距离矢量中的任一个矢量, 其来源都是相同的, 具有同样的统计性, 在样本数目足够大时可以不作变量选择与优化而直接进行有关统计分析。然而, QSAR 研究中所选择的实际体系由于多方面的原因(比如化

物难以合成、化合物不稳定引起性质测定误差偏大、经费不足等)造成所包含的分子样本数一般不是很多,体系中每一个分子中所具有的原子类型种类和数目也不完全一致,有时甚至相差很大,致使 MEDV 矢量中的某些描述子统计显著性变差。这时,就必须对这些描述子进行分析选择,剔除统计显著性太差的描述子,才能进行统计分析。

遗憾的是,至今尚没有进行变量选择的标准方法可供选择。进行变量选择的关键问题之一是变量选择的目标函数如何确定?选择不同的目标函数,可能得到不同的变量优选结果。归纳起来,至今有两大类目标函数可供选择,即以模型对内部样本的估计能力和以模型对外部样本的预测能力为目标两类。因此,变量选择的方法总是与模型的质量联系在一起的。很多研究表明,模型质量的评价不仅应当包括模型对内部样本的估计能力,更重要的还要看其对外部样本的预测能力。然而,目前流行的变量选择方法多以模型估计能力为目标,因而选择的优化变量不一定满足模型预测能力的需要。因此,有必要建立以预测能力为目标的变量选择方法。如何评价预测能力一直是一个国际难题,至今没有统一的标准。但有一个结论是可以肯定的,一个高的估计能力和高的稳定性是模型具有良好预测能力的必要条件。我们通过对多个分子体系如烷烃、多环芳烃、药物分子体系等进行的大量计算实践与分析,设计开发了两种以交互检验为基础(目标函数)的变量选择与模型优化方法,一种基于遗传算法和一种基于最佳子集回归的变量选择方法,前一种方法适用于描述子变量较多的化合物体系,而后一种方法则适用于描述变量较少的化合物体系。

应当指出,在进行变量优选之前,为了进行交互检验分析,样本集中每一个描述子变量的取值至少应有两个样本不为零即至少应具有两个非零样本。因此,在进行变量优化与模型构建之前,应先对描述子变量集进行初步的选择。

3.2.1 描述子变量的初步选择

设变量数据集由 n 个样本组成，每个样本分子结构由 r 个描述子表征。为了能进行交互检验分析，应保证每个描述子变量至少具有两个非零样本数。则变量的初步选择过程^[14]如下：

首先删除 n 个样本 r 个描述子组成的数据集中对所有样本取值都为零或都相等的描述子。这些描述子对所有物质分子性质的贡献都为零或相等，不具有分辨性质，因此首先将其删除。对于具体数据集比如结构类似的药物分子数据集，分子多样性往往表现在不同的取代基上，因而这些分子常常缺少某些原子类型，致使这些原子类型所对应的描述子变量值为零，而各个分子具有的相同骨架可能使得某些描述子具有相同的数值。

其次，删除非零值样本数少的描述子。这些描述子因统计显著性较低，导致相关模型不稳定。到底少到几个样本？统计性不显著到什么程度？目前似无统一规律可寻。但是随样本数据集大小的变化而变化却是不争的事实。从安全性出发，选择删除的描述子，其非零样本数至少应保证少于样本总数的 $1/10$ 。另一方面，为了能进行最基本的交互检验程序，比如 LOO(leave-one-out)交互检验，进入变量优化选择的每个描述子至少应该具有两个非零样本。

在进行变量删除的规划中，我们开始也考虑了在模型构建之前，删除自相关严重（比如两个描述子的相关系数达到 0.7 或 0.8 以上）的两个描述子之一，问题是优先删除哪一个没有标准方法可供应用。尽管原则上删除哪一个从统计上没有显著性差异，但进行删除时还是要非常小心。这是因为相关系数的大小往往随样本数的变化而发生变化，特别是某些极端情况，比如一个描述子对于所有样本都有非零值，而另一个却只有少数如 2~3 个非零样本，这时这两个描述子之间就可能有很高的相关系数，但实际上这两个描述子可能是不相关的。因此，做这一步删除时一定要非常

小心，必须保证有足够大的样本数和相当高的相关系数（比如0.95以上），否则应该在后续变量优化选择过程中删除。

3.2.2 基于遗传算法的变量选择

该方法在 3.2.1 节的初步删除之后进行。其基本思路是利用遗传算法（GA）的基本原理，首先从描述子集中随机选择指定变量数的多个变量组合构成初始群体，以这些选择的变量组合（或子集）为自变量，通过多元线性回归方法（对线性模型）建立与物质分子性质 / 活性的相关模型 并对模型进行 LOO 交互检验，得到交互检验过程中对物质分子性质的预测与相应实验观测值之间的相关系数（用 q^2 表示）或均方根误差（用 $RMSV$ 表示）为目标函数（在 GA 中称适应度函数）通过 GA 的选择、交叉、变异等算子对变量子集进行操作（参见图 3-2），最后选择各指定变量数下的最高 q^2 或最低 $RMSV$ 的变量组合为优化组合。指定的变量数可从 1 或 2 开始 直到样本数的 1/6 或 1/5（取整数）为止^[15]。选择 1/6 或 1/5 的原因是基于一个经验规则的考虑。一般经验认为，一个好的 QSAR 模型，其样本数至少应为变量数的 5 或 6 倍以上。

上述 GA 方法选择的最优变量组合只是说明在指定变量数下的最优组合，并不是全局或整个数据集的最优组合，因此，还必须以均方根误差对指定变量数作图以观察均方根误差的变化趋势，通过综合分析选择全局的最优变量组合。

对描述子变量较多的体系进行变量优化选择时，采用基于 GA 的方法是比较适用的。尽管基于 GA 的变量选择方法有时不一定能找到全局的最优点，即可能选择不到最佳的变量组合，但对于大样本与大变量数据集来说，从时间与效益方面考虑还是合适的。当然，对于小样本或小变量数据集，用下节介绍的最佳子集回归方法更合适。

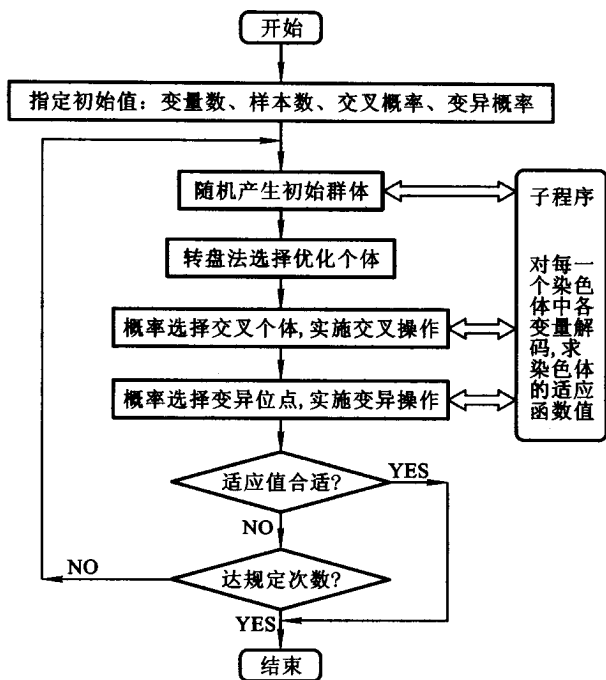


图 3-2 遗传算法选择变量操作示意框图

3.2.3 基于最佳子集回归的变量选择

最佳子集回归 (best subset regression 简称 BSR) 也称所有子集回归 (all subset regression), 它是将变量选择与多元回归结合在一起的最优化方法。与传统 BSR 方法不同的是, 我们不是以模型估计能力为目标函数, 而是以 LOO 交互检验的相关系数 (q) 与均方根误差 (RMSV) 为目标函数来进行最佳子集运算的。以 LOO 交互检验均方根误差为目标函数的变量选择基本过程是: 每次从各种量子集组合中系统地而不是随机地选择一个量子集, 进

行回归建模与交互检验，直到把所有子集都计算完，选择均方根误差最小的子集作为指定变量数下的最佳子集。比如从 20 个原始变量中先选择一个变量分别建立 20 个单变量回归方程，并计算 LOO 相关系数或均方根误差等相关统计量，确定一个变量下的最优结果。然后选择 2 个变量组成的子集，分别建立 190 个两变量回归模型并计算 LOO 过程的有关统计量，确定两个变量下的最优结果；依次运算下去。很明显，当子集中变量数目增加时，其回归方程的组合数按几何级数迅速增加，计算工作量很大。为此，有人提出了采用运筹学方法比如分枝定界算法等来解决计算量的问题。然而，由于我们是以 LOO 交互检验为目标函数的，分枝定界的原理只适用于以估计能力为目标函数的算法，不适用以 LOO 检验为目标的 BSR。为此，我们提出通过设定变量自相关系数阈值与 LOO 交互检验统计量来限制运算次数，从而提高选择速度，建立了命名为 VSMP 的最佳子集变量选择算法（基本过程参见图 3-3）^[16]。运用 VSMP 对每一个指定的变量数选择最佳变量子集，最后以均方根误差对变量数作图，分析确定全局最佳变量子集。

在图 3-3 中， $x(n, m)$ 为自变量数据矩阵，例如我们的 MEDV 矢量描述子集； $y(n)$ 为化合物分子性质 / 生物活性数据矩阵。VSMP 方法运算过程中设定的初值为： m 为 BSR 运行时指定的子集中的变量个数，一般从 1 开始依次增加 1 直到 $n/6$ 或 $n/5$ ； r_{int} 为 BSR 运行时允许的两变量之间的最大相关系数，如果选择的变量子集中任意两个变量之间的相关系数达到或超过这个相关系数时，认为该子集为最差子集，则重新开始选择新子集，以减少运行次数； r_{cn} 与 f_{max} 是两个在 BSR 运行过程中随着优化的进行不断增加的控制参数，以保证下一个选择的子集一步一步向最优结果逼近，同时让较差的子集不继续向下运行从而减少运算次数，进一步提高运行速度。这两个参数的初值，最简单的办法是设定其等于 0。

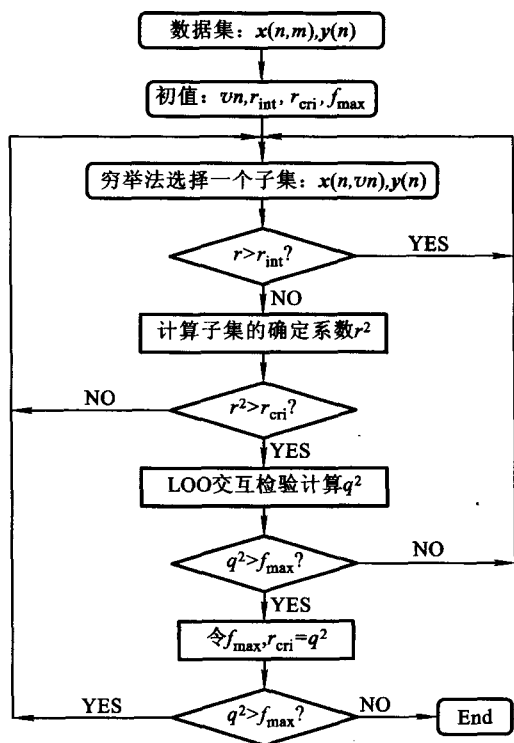


图 3-3 VSMP 法选择变量基本过程框图

3.3 多元线性回归 (MLR)

前节以 LOO 交互检验统计量为目标函数的变量选择与模型化方法不论是遗传算法还是最佳子集回归，都是以线性模型为基础的。研究表明，至今为止，多元线性回归^[17~19] (multiple linear regression 简称 MLR) 方法仍是多变量线性模型研究的最佳方法之一。本节给出 MLR 方法的基本原理与统计参量。

3.3. 模型回归系数

以 n 个样本分子生物活性 / 物理化学性质为因变量 y_i ($i = 1, 2, 3 \dots n$) 各分子描述子为自变量 x_{ij} ($j = 1, 2, 3, \dots, m$)。设存在如下线性关系：

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} + \epsilon_i \\ &= \beta_0 + \sum_{j=1}^m \beta_j x_{ij} + \epsilon_i \end{aligned} \quad (3-3)$$

($i = 1, 2, 3, \dots, n; j = 1, 2, 3, \dots, m$)

式中 n 是样本数, m 是每一个样本分子结构对应的描述子矢量中的描述子个数, x_{ij} 为第 i 个样本第 j 个分子描述子变量, β 为待估计回归系数, ϵ 为误差项。借广义最小二乘拟合可求得 (3-3) 式中各个回归系数 β_j 的估计 b_j 。为方便, 采用矩阵形式求解各回归系数的最小二乘估计^[12]。

将 (3-3) 式表达为两矩阵之乘积, 有:

$$(y_i) = (1 \quad x_{i1} \quad x_{i2} \quad \dots \quad x_{im}) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix} + (\epsilon_i) \quad (3-3a)$$

将 n 个样本分子理化性质或生物活性 y 与分子描述子参数 x 组合起来, 得:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ 1 & x_{31} & x_{32} & \dots & x_{3m} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix} \quad (3-4)$$

以矩阵符号（黑体大写斜体字母表示矩阵）重写上式，有：

$$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times (m+1)} \mathbf{B}_{(m+1) \times 1} + \mathbf{E}_{n \times 1} \quad (3-4a)$$

通过求矩阵广义逆可得到各回归系数的最小二乘估计：

$$\hat{\mathbf{B}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \quad (3-5)$$

将 (3-5) 式代入 (3-4) 式可估计各物质分子（共 n 个）的生物活性 / 理化性质：

$$\hat{\mathbf{Y}} = \mathbf{X} \hat{\mathbf{B}} \quad (3-6)$$

求得回归系数 $\mathbf{B}$ 后，就可利用分子描述子矢量 $\mathbf{x}$ 去预测未知样本分子的生物活性 / 理化性质：

$$\hat{Y}_{un} = \mathbf{x}_{un} \hat{\mathbf{B}} \quad (3-7)$$

这里应当强调指出“计算”、“估计”与“预测”等几个重要概念。如果利用上述回归模型 (3-4) 式去计算用于建立该模型时所用样本集的活性 / 性质时，其计算过程称为“估计”，相应计算结果称为模型估计值。若用模型 (3-4) 式计算并没有参与模型建立的未知（理化性质或生物活性未知）的那些样本活性 / 性质时，其计算过程称为“预测”，相应计算结果称作预测值。这里，估计与预测统称为计算，而估计值与预测值统称为计算值。

另一方面，建立模型所使用的样本集称为“训练集”，其样本称为“内部样本”或“训练集样本”，用于预测的未知样本集称为“外部预测集”，其样本称为“外部预测集样本”。若用于以检验为目的的已知样本集，这些样本不参与建模，称为“外部检验集”，其样本称为“外部检验集样本”。

一个好的定量相关模型必须具有对模型内部样本的良好估计能力，还必须同时具有对外部检验或预测样本的良好预测能力。下一节的内容就是 MLR 方法中用来表征模型估计能力与预测能力的几个重要统计量。

3.3.2 MLR 中几个重要统计量

根据 n 个样本分子的分子描述子 x_{ij} 和实验观测到的生物活性 / 理化性质 y_i (n 个样本的平均值为 $\bar{y}$) 建立起 MLR 多元回归模型, 并由此得到模型的估计值 $\hat{y}_i$ 。利用这些估计值与原实验观测值可以求得几个在模型质量评价中有重要意义的统计量, 如相关系数和均方根误差等。

1. 复相关系数 (R):

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3-8)$$

2 总体标准差 (S_T):

$$S_T^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (y_i - \bar{y})^2 \quad (3-9)$$

3. 剩余标准差 (S_R):

$$S_R^2 = \frac{1}{n-m-1} \cdot \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3-10)$$

4. 解释方差 (EV):

$$\begin{aligned} EV &= 1 - \frac{1}{n-m-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \Big/ \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= 1 - \frac{S_R^2}{S_T^2} \end{aligned} \quad (3-11)$$

5. F 统计量 (F):

$$F = \frac{1}{m} \cdot \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \Big/ \frac{1}{n-m-1} \cdot \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3-12)$$

6. 均方根误差 (RMS):

$$RMS = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3-13)$$

7. 标准误差 (S):

$$S = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3-14)$$

3.3.3 标准回归系数和置信区间

标准回归系数 ($b_{o,j}$) 定义为:

$$b_{o,j} = b_j \cdot \sqrt{\frac{Lxx_j}{Lyy}} \quad (3-15)$$

其中:

$$Lxx_j = \sum_{i=1}^n x_{ij}^2 - \frac{1}{n} \left[\sum_{i=1}^n x_{ij} \right]^2 \quad \text{和} \quad Lyy = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left[\sum_{i=1}^n y_i \right]^2$$

式中 b_j 是由(3-5)式求出的回归系数 β_j 的估计值, x_{ij} 是第 i 个样本分子第 j 个分子描述子变量。

回归系数 β_j 的置信区间为:

$$\beta_j = b_j \pm t_{n-m-1, \alpha} \cdot S_R \cdot \sqrt{C_{jj}} = b_j \pm t_{n-m-1, \alpha} \cdot S_j \quad (3-16)$$

其中 S_j 是回归系数 β_j 的标准误差, 反映回归系数 β_j 的显著性。
 C_{jj} 是矩阵 $(\mathbf{X}^T \mathbf{X})^{-1}$ 中的对角线元素值。

一般来说, 在报告回归方程的同时, 应给出回归系数置信区间或回归系数标准偏差。

3.4 主成分回归 (PCR)

多元线性回归 (MLR) 方法只适用于分子描述子变量数少于样本分子数的线性模型构建。对于描述子数大于样本数的分子体系, 由于存在矩阵奇异而无法求取矩阵广义逆。此时有两种解决办法, 一是采用前节介绍的变量选择方法比如遗传算法或最佳子

集回归方法先进行变量优化选择,再应用 MLR 方法求解;另一个是采用基于主成分分析的主成分回归 (principal component regression 简称 PCR)^[11, 12, 20~22] 或偏最小二乘回归 (partial least square regression 简称 PLSR)方法^[11, 12, 22~27]。PCR 方法仅对描述子变量构成的 $\mathbf{X}$ 矩阵作主成分分解,其主成分与生物活性 / 理化性质矩阵 $\mathbf{Y}$ 无关,便于分析主成分与原始描述子变量的关系。PLSR 方法则同时对 $\mathbf{X}$ 和 $\mathbf{Y}$ 两个数据矩阵都进行主成分分解,并将分解之得分矢量进行线性相关,因而主成分不仅与 $\mathbf{X}$ 有关,而且与 $\mathbf{Y}$ 有关。通常 PLSR 所得的最优主成分数比 PCR 法要小。本节介绍 PCR 方法的基本原理。

3.4.1 PCR 基本原理

设已对分子描述子矩阵 $\mathbf{X}$ 与分子生物活性 / 理化性质 $\mathbf{Y}$ 进行了标准化处理, $\mathbf{Y}$ 与 $\mathbf{X}$ 的相关关系为线性模型:

$$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times m} \mathbf{B}_{m \times 1} + \mathbf{E}_{n \times 1} \quad (3-4a)$$

利用非线性迭代偏最小二乘算法对 $\mathbf{X}$ 矩阵作主成分分解,得两个小矩阵 $\mathbf{T}$ 和 $\mathbf{V}$:

$$\mathbf{X}_{n \times m} = \mathbf{T}_{n \times A} \mathbf{V}_{A \times m}^T + \mathbf{E}_{n \times m} \quad (3-17)$$

式中 $\mathbf{T}$ 称得分矩阵, $\mathbf{V}$ 称载荷矩阵, A 为最佳主成分数。

既然 $\mathbf{Y}$ 与 $\mathbf{X}$ 为线性关系,结合 (3-17) 式,那么 $\mathbf{Y}$ 与 $\mathbf{T}$ 也应是线性关系,有:

$$\mathbf{Y}_{n \times 1} = \mathbf{T}_{n \times A} \mathbf{G}_{A \times 1} + \mathbf{E}_{n \times 1} \quad (3-18)$$

对 (3-18) 式作最小二乘拟合,可计算求得回归系数 $\mathbf{G}$ 矩阵或 $\mathbf{G}$ 矢量的估计:

$$\hat{\mathbf{G}}_{A \times 1} = (\mathbf{T}_{A \times n}^T \mathbf{T}_{n \times A})^{-1} \mathbf{T}_{A \times n}^T \mathbf{Y}_{n \times 1} \quad (3-19)$$

根据 (3-19) 式,利用 (3-18) 式对内部训练集样本分子性质

进行估计或对外部检验集样本分子与未知样本进行预测时，先以训练集样本进行主成分分解求得的正交载荷矩阵 $\mathbf{V}$ 求解待估计样本或预测样本分子描述子集 $\mathbf{X}_{un}$ 的得分矩阵 $\mathbf{T}_{un}$ ，然后再以 (3-18) 式估计或预测分子活性 $\mathbf{Y}_{un}$ ：

$$\hat{\mathbf{T}}_{un \times A} = \tilde{\mathbf{X}}_{un \times m} \mathbf{V}_{m \times A} \quad (3-20)$$

$$\hat{\mathbf{Y}}_{un \times 1} = \hat{\mathbf{T}}_{un \times A} \hat{\mathbf{G}}_{A \times 1} \quad (3-21)$$

原样本分子描述子或称原始变量对各个主成分的贡献可通过载荷矩阵 $\mathbf{V}$ 中各载荷矢量分别对主成分作图进行定性解释，也可以再在得分矩阵 $\mathbf{T}$ 与原变量矩阵 $\mathbf{X}$ 之间作回归处理而进行定量解析。通过每个主成分的载荷矢量对各原始描述子变量作图进行所谓的载荷分析，可以确定对各主成分具有重要影响的描述子变量，删除某些对最佳主成分没有影响或者影响很小的描述子。

主成分回归模型的质量也可由几个统计量进行评价。仿照 MLR 方法，利用实验观测到的训练集样本生物活性 / 理化性质数据 (y_i) (n 个样本的平均活性为 $\bar{y}$) 及由 PCR 模型估计与预测的样本分子生物活性 / 理化性质 ($\hat{y}_i$)，按照公式 (3-8) 到 (3-14) 式等计算主成分回归模型的有关统计量，比如相关系数 (R) 和均方根误差 (RMS) 或标准误差等，从而对 PCR 模型进行评价分析。

3.4.2 从 PCR 模型到 MLR 模型

由 PCR 方法得到的回归模型是一个隐含线性模型，但是各个主成分之载荷矢量是原始分子描述子变量的某种线性组合。据此我们采用矩阵变换技术，将 PCR 隐含模型变换到原始分子描述子构成的类似 MLR 的即所谓习用回归方程或似回归方程 (MLR-like)，从而将主成分空间变换到原始描述子空间。

将 (3-21) 式应用到原始训练集样本生物活性 / 理化性质 ($\mathbf{Y}$) 的估计，并利用 (3-17) 式，有：

$$\hat{Y}_{n \times 1} = T_{n \times A} \hat{G}_{A \times 1} = X_{n \times m} V_{m \times A} \hat{G}_{A \times 1} = X_{n \times m} \hat{B}_{m \times 1} \quad (3-22)$$

其中：

$$\hat{B}_{m \times 1} = V_{m \times A} \hat{G}_{A \times 1} \quad (3-23)$$

因此对应原分子描述子集 X 的习用回归方程为：

$$\hat{Y}_{n \times 1} = X_{n \times m} \hat{B}_{m \times 1} \quad (3-22)$$

3.5 偏最小二乘回归 (PLSR)

3.5.1 PLSR 基本原理

偏最小二乘回归 PLSR 相对 PCR 更进一步，它利用非线性迭代偏最小二乘算法同时将描述子变量矩阵 X 和分子生物活性 / 理化性质矩阵 Y 进行主成分分解，获得描述子矩阵与活性 / 性质矩阵的得分矩阵与载荷矩阵，即

$$X_{n \times m} = T_{n \times A} V_{A \times m}^T + E_{n \times m} \quad (3-17)$$

$$Y_{n \times 1} = R_{n \times A} Q_{A \times 1}^T + F_{n \times 1} \quad (3-24)$$

这里 $E_{n \times m}$, $F_{n \times 1}$ 分别是分子描述子数据矩阵和分子生物活性 / 理化性质数据矩阵的实验误差。由 X 阵得到的 T 矩阵和由 Y 阵得到的 R 矩阵，代表了除去大部分噪声后的响应矩阵与浓度矩阵的浓度信息。因而 T 阵和 R 阵之间应有确定性的关系（线性关系）。即

$$r_i = b_i \cdot t_i \quad (3-25)$$

式中 r_i , t_i 分别取自矩阵 R 和 T 是 R 矩阵和 T 矩阵中的第 i 列。

实际上 PLSR 算法是由两个主成分分析步骤和一个回归步骤组成的。类似于主成分回归，在做主成分分解前一般应对数据矩

阵进行必要的预处理，比如按列均值中心化处理或标准化处理。然而，这样互相独立地进行主成分分解求出的 t 和 r 仍与 PCR 法没有本质区别。为了使由 X 得出的 T 矩阵能与由 Y 得出的 R 矩阵之间有良好的线性关系，因此在 PLSR 算法中，让 X 分解为 T 时引入有关 R 的信息 或在从 Y 分解出 R 时引入 T 的信息 这可以通过在迭代过程中交换迭代变量加以实现。这样两个独立的分解过程就合二为一了。那么一个完整的计算各个主成分下的得分矢量(t 与 r)、载荷矢量(v 与 q) 以及权矢量 w)的 PLSR 算法过程如下：

1 建立 PLSR 模型

对于分子描述子 X 矩阵：

(1) 取 Y 中的某一系列作 r 的迭代初值

(2) 计算权矢量 $w^T: w^T = r^T X / r^T r$

(3) 归一化权矢量 $w^T: w_{new}^T = w_{old}^T / \| w_{old}^T \parallel$

(4) 计算 $t: t = Xw / w^T w$

对于单因变量分子生物活性 / 性质 Y 矩阵 (实际为一矢量)：

(5) 直接令 $q^T = 1$ (因为 Y 为单因变量的矢量)

计算 X 的载荷矢量，并重新标准化得分矢量及权重矢量：

(6) 计算 X 的载荷矢量 $v^T: v^T = t^T X / t^T t$

(7) 归一化 $v^T: v_{new}^T = v_{old}^T / \| v_{old}^T \parallel$

(8) 标准化 X 的得分矢量 $t: t_{new} = t_{old} / \| v_{old}^T \parallel$

(9) 标准化权重矢量 $w_{new}^T: w_{new}^T = w_{old}^T / \| v_{old}^T \parallel$

计算回归系数 b 以用于内部关联：

(10) 计算 $b: b = r^T t / t^T t$

对于第 h 个主成分计算残差：

(11) $X = X - t_h \cdot v_h^T$

(12) $Y = Y - b_h \cdot t \cdot q_h^T$

比较前后两次残差 X 或 Y 是否相等，如果是，模型建立过程

结束 否则 返回第 1)步进行下一个主成分的计算。

2. 未知样本预测

(1) 将未知样本描述子数据矩阵 X_{un} 按模型建立过程中的方法进行预处理或标准化

(2) 令 $h = 0$

(3) 设 $h = h + 1$ 并计算：

$$\begin{aligned}t_h &= X_{un} \cdot w_h^T \\Y_{un} &= Y_{un} + b_h \cdot t_h \cdot q_h^T \\X_{un} &= X_{un} - t_h \cdot v^T\end{aligned}$$

(4) 若 $h > k$ 转第 5 步 否则返回第 3 步

(5) 上述计算过程计算得到的 Y 是标准化的结果，需按标准化或预处理步骤中的相反过程操作，以恢复到原始数据坐标（即分子生物活性 / 理化性质）

3.5.2 从 PLSR 模型到 MLR 模型

与 PCR 方法一样 由 PLSR 方法得到的回归模型是一个隐含线性模型，一个关于主成分的隐式模型，不是关于原始描述子变量的显式线性模型。但是由 PLSR 方法通过主成分分解的各个主成分的各载荷矢量都是原始分子描述子变量的某种线性组合。因此，通过某些矩阵变换或操作，应该能够将 PLSR 关于主成分的隐含模型变换为关于原始分子描述子变量的显式模型，这个模型称为习用回归方程或似回归方程（MLR-like）^[28]。利用似回归方程 就可以像 MLR 方法一样使用 PLSR 模型的结果，直接将未知样本的分子描述子结果代入似回归模型来计算该分子的生物活性 / 理化性质。

已知 PLSR 模型的最佳主成分数为 A ，各最佳主成分对应的得分矢量、载荷矢量、权矢量以及得分矢量的回归系数分别为 t, w, r ，那么根据 X 矩阵和 Y 矩阵与 t 以及 r 之间的关系，可以

得到似回归方程对应的 MLR 形式的系数矩阵 $\mathbf{B}$:

定义新的权向量 $\mathbf{w}^*$ 或矩阵 $\mathbf{W}^{*[29]}$:

$$\mathbf{W}^* = \mathbf{W}(\mathbf{V}^T \mathbf{W})^{-1} \quad (3-26)$$

由于

$$t_{ia} = \sum_{k=1}^m w_{ka}^* \cdot x_{ik} \quad (\mathbf{T}_{n \times A} = \mathbf{X}_{n \times m} \cdot \mathbf{W}_{m \times A}^*)$$

则

$$\begin{aligned} y_{i1} &= \sum_{a=1}^A r_a \cdot t_{ia} + e_{i1} \quad (\mathbf{Y}_{n \times 1} = \mathbf{T}_{n \times A} \cdot \mathbf{R}_{A \times 1}^T + \mathbf{E}_{n \times 1}) \\ &= \sum_{a=1}^A r_a \cdot \sum_{k=1}^m w_{ka}^* \cdot x_{ik} + e_{i1} \\ &(\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times m} \cdot \mathbf{W}_{m \times A}^* \cdot \mathbf{R}_{A \times 1}^T + \mathbf{E}_{n \times 1}) \end{aligned}$$

与习用回归方程 (3-4a) 式 :

$$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times m} \mathbf{B}_{m \times 1} + \mathbf{E}_{n \times 1} \quad (3-4a)$$

相比可知

$$\mathbf{B}_{m \times 1} = \mathbf{W}_{m \times A}^* \mathbf{R}_{A \times 1}^T \quad (3-27)$$

将 (3-26) 式代入 (3-27) 式, 得关于原始描述子变量的习用回归方程的回归系数 :

$$\mathbf{B}_{m \times 1} = \mathbf{W}_{m \times A} (\mathbf{V}_{A \times m}^T \cdot \mathbf{W}_{m \times A})^{-1} \cdot \mathbf{R}_{A \times 1}^T \quad (3-28)$$

3.5.3 原始描述子变量的重要性分析

尽管 PLSR 模型是关于主成分的隐含模型, 但是各主成分都是原始变量的线性组合, 不同原始变量对不同主成分的贡献是不一样的, 因而对分子生物活性 / 理化性质 ($\mathbf{Y}$) 影响的重要性也是不一样的。原始变量的投影重要性指标 (variable importance in pro-

jection 简称 VIP)^[29]可用来分析原始变量对 $\mathbf{Y}$ 影响的重要性, 变量的 VIP 值越大, 该变量的重要性也就越大。

定义 $\mathbf{Y}$ 矩阵 (对于单因变量实际上是一矢量) 与第 a 个主成分的得分矢量 $\mathbf{t}_a$ 之间的相关系数 $r(\mathbf{y}, \mathbf{t}_a)$ 为:

$$r^2(\mathbf{y}, \mathbf{t}_a) = 1 - \frac{\sum_{i=1}^n (y_i - t_{ia})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3-29)$$

则第 j 个变量的投影重要性指标 VIP_j 计算如下:

$$VIP_j = \text{sqr} \left[m \cdot \sum_{a=1}^A w_{ja}^2 \cdot r^2(\mathbf{y}, \mathbf{t}_a) / \sum_{a=1}^A r^2(\mathbf{y}, \mathbf{t}_a) \right] \quad (3-30)$$

3.6 模型质量评价

如何评价 QSAR 模型的质量是模型能否有效应用的前提。一个定量相关模型的质量评价至少应该包括模型对内部样本的估计能力、模型稳定性及对外部样本的预测能力等三个方面。本节对模型质量评价的一般原则做简要说明。

3.6.1 模型估计能力

模型估计能力通过内部样本建模过程中获得的几个重要统计量进行评价。设 n 个内部样本分子的因变量 (如分子生物活性 / 理化性质) 观测值为 y_i (n 个样本的平均值为 $\bar{y}$) 由 MLR 模型得到的因变量估计值为 $\hat{y}$ 。那么模型估计能力一般由 (复) 相关系数、均方根误差以及回归系数的标准偏差进行评价。一般说, 相关系数越接近于 1, 均方根误差与回归系数标准偏差越小, 模型对内部样本的估计能力就越强。研究表明, 模型估计能力评价中最重要的统计量是 (估计) 均方根误差 $RMSE$, 其次是复相关系数 R 。进

入 QSAR 模型中的分子描述子选择是很重要的影响因素，因此在评价模型估计能力时，应该考虑各个分子描述子对应回归系数的标准偏差大小。

3.6.2 模型稳定性

模型稳定性是模型具有良好预测能力的前提。是一个比模型估计能力更重要的质量评价指标，通常通过交互检验 (cross-validation 简称 CV) 程序来完成。在交互检验中有一种被广泛采用的方法就是 LOO 交互检验法。LOO 是英文 leave-one-out 中三个首字母的缩写，其实验步骤（对于含 n 个样本的数据集）为：先抽出第 1 个样本作为外部检验样本，余下的 $n-1$ 个样本作为训练集建立回归模型，并用这个模型去预测抽出的作为外部检验集样本的因变量值；然后将第 1 个样本放回原样本数据集，依次抽出第 2 个样本作为外部检验集样本，同样以余下的 $n-1$ 个样本为训练集建立回归模型，并预测第 2 个样本的因变量值；将第 2 个样本放回原数据集，依次按照“抽出一个样本→余下样本建模→预测抽出样本→放回抽出样本”的顺序对原样本集进行操作，直到所有样本都被抽出一一次并进行预测时为止。

完成 LOO 交互实验后，通过最小二乘法计算 n 次抽出样本的因变量 LOO 预测值 (y_i^{LOO}) 与原抽出样本的实验因变量值 (y_i) 之间的 (检验) 相关系数 q 与 LOO 检验均方根误差 (RMSV) 等统计量以评价模型稳定性。

也有人采用随机分组或定制分组或每次抽出 N 个样本的所谓 LNO 交互检验法等对模型稳定性进行评价。随机抽样方法以随机方式从样本集中抽出部分样本作为外部检验集，以未抽出的样本建立模型，用所建模型预测检验集样本分子。为保证评价的可靠性，必须进行多次随机抽样实验方能证明模型的稳定性或预测能力。该法缺点是重复性不易控制。定制分组方法是一种简便但不甚完全的检验方法。一般以样本集前面的样本分子作为训练

集建立相关模型，然后预测后面余下样本分子的活性。当然也可反方向进行。LNO 法过程如下：先将样本集分为若干组比如 g 组 再进行类似 LOO 法的“抽出一组样本→余下样本建模→预测抽出样本”的操作，直到每一组样本被抽出一次并进行预测为止。不同的仅是每次抽出的是一组样本而不是一个样本。最后计算所有样本因变量的 LOO 预测值与原实验观测值之间的相关系数 (q) 与均方根误差 ($RMSV$) 等有关统计参量，以评价模型的稳定性。

3.6.3 模型预测能力

具有高交互检验相关系数或低均方根误差是模型具有良好预测能力的前提，要证明模型预测能力还必须应用外部检验样本。应用训练集样本建立的回归模型预测外部样本的生物活性（以 y_i^p 表示）并计算预测值 (y_i^p) 与原实验观测值之间的预测相关系数 (r_p) 和预测均方根误差 ($RMSP$)。相关系数越高，均方根误差越低，模型对外部样本的预测能力就越好。

从目前的研究来看，一个模型的估计能力、稳定性与预测能力之间有何联系尚没有统一的判别标准。然而，一般来说，一个模型如果没有好的估计能力，就不可能有好的稳定性和预测能力；模型稳定性是良好预测的必要条件。因此，在报告 QSAR 模型时 应该给出有关估计和稳定性的统计量，同时也应该给出预测能力及其所使用的检验集或预测集的有关信息。

3.7 应用程序 VSMP 简介

计算机应用程序的编写与开发是发展与应用分子描述子与多元统计技术的最重要也最关键的技术之一。在 QSAR 研究中 如果不能开发有关应用程序，就不可能发展新型分子描述子，也不可能改进现有描述子，更不可能开发应用新的统计技术。作者在发

展分子电性距离矢量方法进行分子结构表征和构效关系研究中，集成与开发编制了有关计算机应用程序，以期实现我们自行设计的分子电性距离矢量计算并进行以交互检验而不仅仅是模型估计为目标函数的 QSAR 研究目标，建立了分子描述子计算、变量优化选择与模型化方法有机结合的应用程序平台，暂且称为 VSMP，即基于预测的变量选择与模型化方法 (variable selection and modeling based on the prediction)。应用程序 VSMP 是在 Visual BASIC 面向对象程序语言平台上进行开发而建立的，它是一个自封闭且易嵌入新的程序模块的 QSAR 研究系统，目前包括 MEDV 描述子和原子类型电拓扑状态指数等分子描述子的计算、描述子变量的初步选择、生物活性变换、基于 LOO 检验预测的最佳子集回归算法、主成分回归以及偏最小二乘回归算法等几个程序模块。

3.7.1 VSMP 1.0 主要模块与功能

VSMP 1.0 主要包括矢量分子描述子、样本性质与数据集、变量选择与模型化、在线帮助等主要模块，分别完成分子描述子计算与预处理、数据集操作、最佳子集回归、主成分回归和偏最小二乘回归等任务。

在矢量分子描述子模块中，根据分子中各非氢原子的原子类型或原子属性分类方案，以及各非氢原子之间的连接关系，输入分子的拓扑结构与电性信息，计算每个分子的距离矩阵与邻接矩阵，进而计算分子电性距离矢量 (MEDV - 13) 和原子类型电拓扑状态指数 (ETSI) 等两大类分子描述子。在该模块中的另一个重要功能是对计算的 MEDV 和 ETSI 原始分子描述子进行初步处理，比如删除非零样本数为 0 或很少的描述子，因为这些描述子对分子结构的分辨率没有多少贡献，同时在进一步的 QSAR 分析中统计显著性很差，因此通常应当优先删除。初步处理还包括描述子变量之间的自相关系数计算、样本相似性分析以及描述子的选择等。

样本性质与数据集模块包括样本性质的数据变换、数据集的

构建以及训练集选择等子模块。在 QSAR 研究中，实验测定的样本性质有时不能满足在宽范围内分布均匀的要求，因此常常需要进行预处理比如对数变换等。VSMP 可以进行平移、倒数、常用对数与自然对数、开方等预处理工作。如何选择训练集是 QSAR 研究中的重要关键问题之一，本模块可实现 2 种选择训练集的方法，即直接以前面部分的样本作为训练集，后面部分的样本为检验集以及先按样本性质大小进行排序，而后从排序数据集中均匀地抽出若干样本为训练集，余下的为检验集。

变量选择与模型化模块包括三个部分：最佳子集、主成分回归和偏最小二乘回归。在最佳子集回归子模块，① 可对整体数据集或训练集样本，按给定的优化变量个数分别计算以 LOO 预测均方根误差最小为目标函数的优化描述子变量组合（优化变量子集），并作出均方根误差随变量数的变化趋势图，最终结合各变量的回归系数与相应标准偏差分析，确定最佳变量组合（最佳子集）；

应用多元线性回归技术建立最佳变量子集的 QSAR 模型 输出模型估计的分子性质、实验性质及其误差，给出 MLR 模型的各个回归系数及相应的标准偏差以及标准回归系数，给出估计或校正的相关系数、均方根误差等重要统计量及 LOO 预测或校验过程中的相应统计量；③ 应用训练集构建的模型对检验集或对未知样本进行预测，从而完成模型检验过程。在主成分回归子模块，应用整体数据集或训练集进行主成分分析，输出各个主成分下校正与 LOO 预测的相关系数和均方根误差等重要统计量，作均方根误差随主成分数的变化趋势图，从而确定最佳主成分；② 建立最佳主成分下的 PCR 模型，输出相应似回归方程的各个回归系数和几个重要统计量；③ 以校正集 PCR 模型预测检验集或未知样本的性质，完成模型校验过程。在偏最小二乘回归子模块，① 应用整体数据集或训练集进行偏最小二乘分析，输出各个主成分下校正与 LOO 预测的相关系数和均方根误差等重要统计量，作均方根误差随主成分数的变化趋势图，从而确定最佳主成分；② 建立最佳

主成分下的偏最小二乘回归模型，输出相应似回归方程的各个回归系数和几个重要统计量；同时计算各个变量的 VIP 函数并进行重要性分析；以校正集偏最小二乘回归模型预测检验集样本的性质 / 活性，或预测未知样本的性质，完成模型校验过程。

完成上述模块功能的 VSMP 1.0 应用程序的主菜单设计如图 3-4 所示。

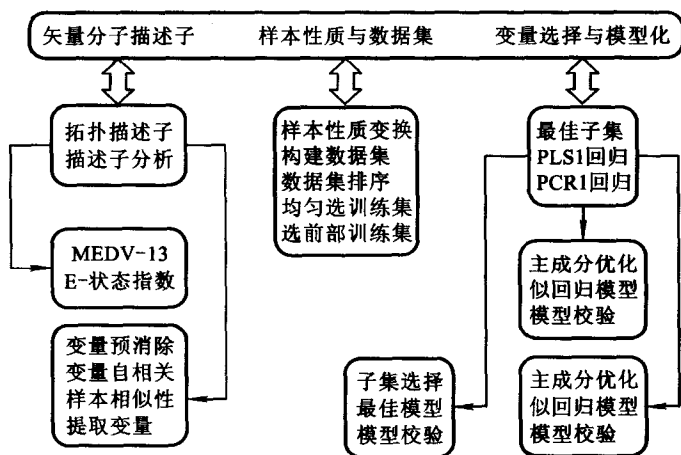


图 3-4 应用程序 VSMP 1.0 主要功能菜单示意图

3.7.2 应用程序 VSMP 的工作流程

应用程序 VSMP 的基本工作流程如图 3-5 所示。它与其他相关程序的不同之处在于：

采用原子属性编号输入分子结构中的子结构信息如电拓扑信息或化学键特性信息，以原子连接关系输入分子拓扑结构；

能计算分子电性距离矢量和电拓扑状态指数；

能以 LOO 检验预测为目标函数，应用最佳子集回归程序优化描述子变量的选择；

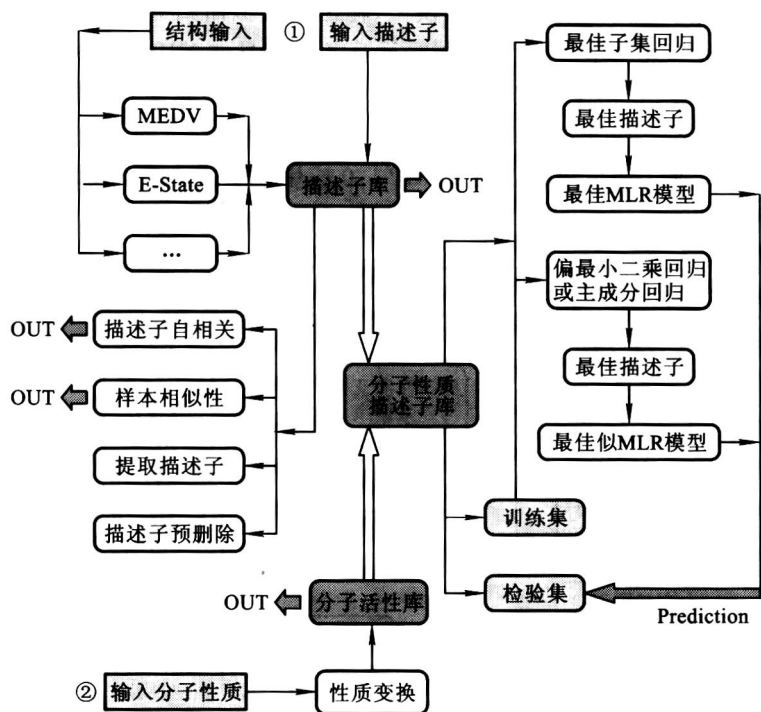


图 3-5 应用程序 VSMP 基本工作流程示意图

能以 LOO 检验预测为目标函数，应用主成分回归和偏最小二乘回归程序获取各个主成分的统计信息，进而建立最佳主成分条件下的似 MLR 模型，并进行原始变量分析。

参 考 文 献

- [1] Randic M. On characterization of molecular Branching. J. Am. Chem. Soc., 1975, 97: 6609~6615
- [2] Randic M. On molecular identification numbers. J. Chem. Inf. Comput.

Sci., 1984, 24: 164~175

- [3] Randic M. Generalized molecular descriptors. J. Math. Chem., 1991, 7: 155~168
- [4] Randic M.* Novel molecular description for structure - property studies. Chem. Phys. Lett., 1993, 211: 478~483
- [5] Randic M. Molecular shape profiles. J. Chem. Inf. Comput. Sci., 1995, 35: 373~382
- [6] Balban A T. High discrimination distance - based topological index. Chem. Phys. Lett., 1982, 89: 399~404
- [7] Estrada E. Edge adjacency relationships and a novel topological index related to molecular volume. J. Chem. Inf. Comput. Sci., 1995, 35: 31~33
- [8] Gutman I. Selected properties of the Schultz molecular topological index. J. Chem. Inf. Comput. Sci., 1994, 34: 1087~1089
- [9] Hosoya H. Topological index. A newly proposed quantity characterising the topological nature of structural isomers of saturated hydrocarbons. Bull. Chem. Soc. Japan, 1971, 44: 2332~2339
- [10] Mohar B, Babic D, Trinajstić N. A novel definition of the Wiener index for trees. J. Chem. Inf. Comput. Sci., 1993, 33: 153~154
- [11] 俞汝勤. 化学计量学导论. 长沙: 湖南教育出版社, 1991
- [12] 刘树深, 易忠胜. 基础化学计量学. 北京: 科学出版社, 1999
- [13] Liu S S, Cai S X, Cao C Z, Li Z L. Molecular electronegative distance vector (MEDV) relating to 15 properties of alkanes. J. Chem. Inf. Comput. Sci., 2000, 40(6): 1337~1348
- [14] Liu S S, Yin C S, Cai S X, Li Z L. QSAR study of steroid benchmark and dipeptides based on MEDV - 13. J. Chem. Inf. Comput. Sci., 2001, 41(2): 321~329
- [15] Liu S S, Yin C S, Wang X D, Wang L S. QSAR studies on dipeptides based on a combinatorial MHDV - GA - MLR method. J. Chin. Chem. Soc., 2002, 49(6): 1089~1096
- [16] Liu S S, Liu H L, Yin C S, Wang L S. VSMP: a novel variable selection and modeling method based on the prediction. J. Chem. Inf. Comput. Sci., 2003, 43(3): 964~969

- [17] 周纪芄. 回归分析. 上海: 华东师范大学出版社, 1993
- [18] 张小蒂. 应用回归分析. 杭州: 浙江大学出版社, 1991
- [19] 任若恩, 王惠文. 多元统计数据分析(理论、方法、实例). 北京: 国防工业出版社, 1997
- [20] Susnow R, Schutt C, Rabitz H. Principal component analysis of dipeptides. *J. Comput Chem.*, 1994, 15: 963~980
- [21] Sharaf M A, Illman D L, Kowalski B R. *Chemometrics*. New York: John Wiley & Sons, 1986
- [22] Geladi P, Kowalski B R. Partial least - square regression: a tutorial. *Anal. Chim. Acta*, 1986, 185: 1~17
- [23] Haaland D M, Thomas E V. PLS methods for spectral analyses 1. Relation to the other quantitative calibration methods and the extraction of qualitative information. *Anal. Chem.*, 1988, 60: 1193~1201
- [24] Baroni M, Clementi S, Cruciani G, et al. Predictive ability of regression models. Part 2. Selection of the best predictive PLS model. *J. Chemom.*, 1992, 6: 347~356
- [25] Cruciani G, Watson K A. Comparative molecular field analysis using GRID force field and GOLPE variable selection methods in a study of inhibitors of glycogen. *J. Med. Chem.*, 1994, 37: 2589~2601
- [26] Wakeling I N, Macfie H J H. A robust PLS procedure. *J. Chemom.*, 1992, 6: 189~198
- [27] 许禄. 化学计量学方法. 北京: 科学出版社, 1995
- [28] Wold S, Sjostrom M, Eriksson L. PLS - regression: a basic tool of chemometrics. *Chemom. Intel. Lab. Syst.* 2001. 58(2): 109~130
- [29] 王惠文. 偏最小二乘回归方法及其应用. 北京: 国防工业出版社, 1999

第 4 章

MEDV - 4 用于 QSPR 研究

第二章提出了表征有机物（如药物与污染物）分子结构的新型矢量描述子即分子电性距离矢量（MEDV），系统地讨论了建立 MEDV 矢量描述子的基本方法。本章将应用第三章的矢量描述子研究方法，以有机化合物的理化性质为实例，通过定量结构 - 性质相关（QSPR）模型的建立与检验，证明基于 4 种原子类型的电性距离矢量即 MEDV - 4 用于表征简单有机分子结构的有效性和适用性。

4.1 烷烃分子结构表征与沸点 QSPR 模型

烷烃分子是研究分子描述子的基本对象。很多分子描述子比如业已广泛应用的 Wiener 指数、Randic 指数、Hosoya 指数等各种拓扑描述子^[1~4]多是从烷烃分子入手或涉及烷烃分子的。这可能是因为烷烃分子理化性质数据来源丰富，实验测定准确度高以及可供应用的分异构体数量多。因此，本章首先以烷烃分子为例，探讨 MEDV - 4 矢量分子描述子在烷烃各种理化性质 QSPR 研究中的应用。

4.1.1 烷烃分子的 MEDV-4

烷烃分子仅由碳氢两种元素原子构成。非氢原子只有碳原子，因而所有非氢原子相对电负性都等于 1。烷烃分子骨架中所有化学键都为碳碳共价单键，相对键长都等于 1。烷烃分子涉及 4 种原子类型，但只涉及 1 种原子属性即“C1”。因此，对于烷烃分子，其 MEDV-4 描述子的计算具有相当简单的形式^[5]：

$$x_r = m_{kl} = \sum_{i \in k, j \in l} \frac{1}{d_{ij}^2} \quad (k = 1, 2, 3, 4; k \leq l \leq 4) \quad (2-3)$$

由上式可知，MEDV-4 主要从两方面表征烷烃分子结构。4 种原子类型相互结合表征了烷烃分子中的各种结构包括支化结构等。如第 1 原子类型 ($k=1$) 表征分子两端的碳原子，第 2 原子类型表征分子中间的直链结构，而第 3 原子类型则表征带一个支链的碳原子，第 4 原子类型则指带两个支链的碳原子。计算公式中相对距离 d_{ij} （在烷烃中也是拓扑距离）解释了各种原子类型表征的各种支化结构中非氢原子之间的相互连接关系，即分子拓扑结构信息。由于 MEDV-4 所依据的是以原子类型为基础的矢量分子描述子，不像有关只具有标量性质的拓扑指数，不会丢失相关的拓扑信息，因而具有高分辨率。对于所研究的从含 1 个碳原子到 10 个碳原子的所有 150 个烷烃分子异构体而言，MEDV-4 对它们的分子结构具有完全的分辨率 ($R_D=1$)。这 150 个烷烃的名称及 10 个 MEDV-4 分子描述子的取值分别列入表 4-1A 和表 4-1B 中。

为了比较，引入选择性系数^[6]即针对相同碳原子数的多种同分异构体定义的分辨率，并给出几个有关拓扑指数与 MEDV-4 选择性系数结果（表 4-2）。由表可知，MEDV-4 在这几种常用拓扑描述子中对异构体的分辨率最高。

表 4-1A 150 个烷烃分子的序号与名称

No	Alkane	No	Alkane	No	Alkane
1	methane	14	heptane	27	2,2 - dimethylhexane
2	ethane	15	2 - methylhexane	28	2,3 - dimethylhexane
3	propane	16	3 - methylhexane	29	2,4 - dimethylhexane
4	butane	17	2,2 - dimethylpentane	30	2,5 - dimethylhexane
5	2 - methylpropane	18	2,3 - dimethylpentane	31	3,3 - dimethylhexane
6	pentane	19	2,4 - dimethylpentane	32	3,4 - dimethylhexane
7	2 - methylbutane	20	3,3 - dimethylpentane	33	3 - ethylhexane
8	2,2 - dimethylpropane	21	3 - ethylpentane	34	2,2,3 - trimethylpentane
9	hexane	22	2,2,3 - trimethylbutane	35	2,2,4 - trimethylpentane
10	2 - methylpentane	23	octane	36	2,3,3 - trimethylpentane
11	3 - methylpentane	24	2 - methylheptane	37	2,3,4 - trimethylpentane
12	2,2 - dimethylbutane	25	3 - methylheptane	38	2 - methyl - 3 - ethylpentane
13	2,3 - dimethylbutane	26	4 - methylheptane	39	3 - methyl - 3 - ethylpentane

续表

No	Alkane	No	Alkane	No	Alkane
40	2, 2, 3, 3 - tetramethylbutane	53	4, 4 - dimethylheptane	66	3 - methyl - 3 - ethylhexane
41	nonane	54	3 - ethylheptane	67	3 - methyl - 4 - ethylhexane
42	2 - methyloctane	55	4 - ethylheptane	68	2, 2, 3, 3 - tetramethylpentane
43	3 - methyloctane	56	2, 2, 3 - trimethylhexane	69	2, 2, 3, 4 - tetramethylpentane
44	4 - methyloctane	57	2, 2, 4 - trimethylhexane	70	2, 2, 4, 4 - tetramethylpentane
45	2, 2 - dimethylheptane	58	2, 2, 5 - trimethylhexane	71	2, 3, 3, 4 - tetramethylpentane
46	2, 3 - dimethylheptane	59	2, 3, 3 - trimethylhexane	72	2, 2 - dimethyl - 3 - ethylpentane
47	2, 4 - dimethylheptane	60	2, 3, 4 - trimethylhexane	73	2, 3 - dimethyl - 3 - ethylpentane
48	2, 5 - dimethylheptane	61	2, 3, 5 - trimethylhexane	74	2, 4 - dimethyl - 3 - ethylpentane
49	2, 6 - dimethylheptane	62	2, 4, 4 - trimethylhexane	75	3, 3 - diethylpentane
50	3, 3 - dimethylheptane	63	3, 3, 4 - trimethylhexane	76	decane
51	3, 4 - dimethylheptane	64	2 - methyl - 3 - ethylhexane	77	2 - methylnonane
52	3, 5 - dimethylheptane	65	2 - methyl - 4 - ethylhexane	78	3 - methylnonane

续表

No	Alkane	No	Alkane	No	Alkane
79	4 - methylnonane	92	4, 5 - dimethyloctane	105	2, 4, 6 - trimethylheptane
80	5 - methylnonane	93	3 - ethyloctane	106	2, 5, 5 - trimethylheptane
81	2, 2 - dimethyloctane	94	4 - ethyloctane	107	3, 3, 4 - trimethylheptane
82	2, 3 - dimethyloctane	95	2, 2, 3 - trimethylheptane	108	3, 3, 5 - trimethylheptane
83	2, 4 - dimethyloctane	96	2, 2, 4 - trimethylheptane	109	3, 4, 4 - trimethylheptane
84	2, 5 - dimethyloctane	97	2, 2, 5 - trimethylheptane	110	3, 4, 5 - trimethylheptane
85	2, 6 - dimethyloctane	98	2, 2, 6 - trimethylheptane	111	2 - methyl - 3 - ethylheptane
86	2, 7 - dimethyloctane	99	2, 3, 3 - trimethylheptane	112	2 - methyl - 4 - ethylheptane
87	3, 3 - dimethyloctane	100	2, 3, 4 - trimethylheptane	113	2 - methyl - 5 - ethylheptane
88	3, 4 - dimethyloctane	101	2, 3, 5 - trimethylheptane	114	3 - methyl - 3 - ethylheptane
89	3, 5 - dimethyloctane	102	2, 3, 6 - trimethylheptane	115	3 - methyl - 4 - ethylheptane
90	3, 6 - dimethyloctane	103	2, 4, 4 - trimethylheptane	116	3 - methyl - 5 - ethylheptane
91	4, 4 - dimethyloctane	104	2, 4, 5 - trimethylheptane	117	4 - methyl - 3 - ethylheptane

续表

No	Alkane	No	Alkane	No	Alkane
118	4 - methyl - 4 - ethylheptane	129	2, 3, 4, 4 - tetramethylhexane	140	3, 4 - dimethyl - 3 - ethylhexane
119	4 - propylheptane	130	2, 3, 4, 5 - tetramethylhexane	141	3, 3 - diethylhexane
120	4 - isopropylheptane	131	3, 3, 4, 4 - tetramethylhexane	142	3, 4 - diethylhexane
121	2, 2, 3, 3 - tetramethylhexane	132	2, 2 - dimethyl - 3 - ethylhexane	143	2 - methyl - 3 - isopropylhexane
122	2, 2, 3, 4 - tetramethylhexane	133	2, 2 - dimethyl - 4 - ethylhexane	144	2, 2, 3, 3, 4 - pentamethylpentane
123	2, 2, 3, 5 - tetramethylhexane	134	2, 3 - dimethyl - 3 - ethylhexane	145	2, 2, 3, 4, 4 - pentamethylpentane
124	2, 2, 4, 4 - tetramethylhexane	135	2, 3 - dimethyl - 4 - ethylhexane	146	2, 2, 3 - trimethyl - 3 - ethylpentane
125	2, 2, 4, 5 - tetramethylhexane	136	2, 4 - dimethyl - 3 - ethylhexane	147	2, 2, 4 - trimethyl - 3 - ethylpentane
126	2, 2, 5, 5 - tetramethylhexane	137	2, 4 - dimethyl - 4 - ethylhexane	148	2, 3, 4 - trimethyl - 3 - ethylpentane
127	2, 3, 3, 4 - tetramethylhexane	138	2, 5 - dimethyl - 3 - ethylhexane	149	2 - trimethyl - 3, 3 - diethylpentane
128	2, 3, 3, 5 - tetramethylhexane	139	3, 3 - dimethyl - 4 - ethylhexane	150	2, 4 - ditrimethyl - 3 - isopropylpentane

表 4-1B 150 个烷烃分子的 MEDV-4 描述子值

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.2500	2.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4	0.1111	2.5000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	0.7500	0.0000	3.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	0.0625	2.7222	0.0000	0.0000	2.2500	0.0000	0.0000	0.0000	0.0000	0.0000
7	0.4722	1.5000	2.2500	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
8	1.5000	0.0000	0.0000	4.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	0.0400	2.8472	0.0000	0.0000	3.6111	0.0000	0.0000	0.0000	0.0000	0.0000
10	0.3750	1.9722	2.1111	0.0000	1.0000	1.2500	0.0000	0.0000	0.0000	0.0000
11	0.2847	2.7222	1.5000	0.0000	0.2500	2.0000	0.0000	0.0000	0.0000	0.0000
12	1.0833	1.7500	0.0000	3.2500	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
13	0.9444	0.0000	5.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
14	0.0278	2.9272	0.0000	0.0000	5.0347	0.0000	0.0000	0.0000	0.0000	0.0000

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
15	0.3300	2.2083	2.0625	0.0000	2.2500	1.3611	0.0000	0.0000	0.0000	0.0000
16	0.2136	3.0972	1.3611	0.0000	1.3611	2.2500	0.0000	0.0000	0.0000	0.0000
17	0.9375	2.3333	0.0000	3.1111	1.0000	0.0000	1.2500	0.0000	0.0000	0.0000
18	0.7083	1.4722	4.1111	0.0000	0.0000	1.2500	0.0000	1.0000	0.0000	0.0000
19	0.7500	1.0000	4.4444	0.0000	0.0000	2.0000	0.0000	0.2500	0.0000	0.0000
20	0.7569	3.2222	0.0000	2.5000	0.2500	0.0000	2.0000	0.0000	0.0000	0.0000
21	0.1875	3.6667	0.7500	0.0000	0.7500	3.0000	0.0000	0.0000	0.0000	0.0000
22	1.6667	0.0000	2.7500	3.5000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
23	0.0204	2.9828	0.0000	0.0000	6.4983	0.0000	0.0000	0.0000	0.0000	0.0000
24	0.3056	2.3508	2.0400	0.0000	3.6111	1.4236	0.0000	0.0000	0.0000	0.0000
25	0.1789	3.2883	1.3125	0.0000	2.6736	2.3611	0.0000	0.0000	0.0000	0.0000
26	0.1528	3.4272	1.2222	0.0000	2.5347	2.5000	0.0000	0.0000	0.0000	0.0000
27	0.8700	2.6319	0.0000	3.0625	2.2500	0.0000	1.3611	0.0000	0.0000	0.0000
28	0.6147	1.9583	3.9236	0.0000	1.0000	1.6111	0.0000	1.0000	0.0000	0.0000

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
29	0.5661	2.2361	3.6458	0.0000	0.2500	3.1111	0.0000	0.2500	0.0000	0.0000
30	0.6600	1.4444	4.2500	0.0000	1.0000	2.5000	0.0000	0.1111	0.0000	0.0000
31	0.6372	3.7083	0.0000	2.3611	1.3611	0.0000	2.2500	0.0000	0.0000	0.0000
32	0.4983	2.8472	3.2222	0.0000	0.1111	2.5000	0.0000	1.0000	0.0000	0.0000
33	0.1425	3.9444	0.6111	0.0000	1.9722	3.2500	0.0000	0.0000	0.0000	0.0000
34	1.3819	1.5833	2.0000	3.3611	0.0000	1.0000	0.2500	0.0000	1.0000	0.0000
35	1.3750	1.2500	2.3333	3.2222	0.0000	1.0000	1.0000	0.0000	0.2500	0.0000
36	1.2917	1.7222	2.6111	2.7500	0.0000	0.2500	1.0000	0.0000	1.0000	0.0000
37	1.1944	0.0000	6.9444	0.0000	0.0000	0.0000	0.0000	2.2500	0.0000	0.0000
38	0.5625	2.6667	3.2222	0.0000	0.2500	2.5000	0.0000	1.0000	0.0000	0.0000
39	0.5208	4.4167	0.0000	1.7500	0.7500	0.0000	3.0000	0.0000	0.0000	0.0000
40	2.5000	0.0000	0.0000	7.5000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
41	0.0156	3.0236	0.0000	0.0000	7.9897	0.0000	0.0000	0.0000	0.0000	0.0000
42	0.2908	2.4464	2.0278	0.0000	5.0347	1.4636	0.0000	0.0000	0.0000	0.0000

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
43	0.1593	3.4064	1.2900	0.0000	4.0747	2.4236	0.0000	0.0000	0.0000	0.0000
44	0.1229	3.5939	1.1736	0.0000	3.8872	2.6111	0.0000	0.0000	0.0000	0.0000
45	0.8333	2.8144	0.0000	3.0400	3.6111	0.0000	1.4236	0.0000	0.0000	0.0000
46	0.5678	2.2119	3.8525	0.0000	2.2500	1.7847	0.0000	1.0000	0.0000	0.0000
47	0.4931	2.6286	3.4844	0.0000	1.3611	3.4236	0.0000	0.2500	0.0000	0.0000
48	0.4967	2.5869	3.4775	0.0000	1.3611	3.5625	0.0000	0.1111	0.0000	0.0000
49	0.6111	1.6944	4.1600	0.0000	2.2500	2.7222	0.0000	0.0625	0.0000	0.0000
50	0.5800	3.9619	0.0000	2.3125	2.6736	0.0000	2.3611	0.0000	0.0000	0.0000
51	0.4150	3.2883	3.0347	0.0000	1.1736	2.8611	0.0000	1.0000	0.0000	0.0000
52	0.3925	3.4272	2.8472	0.0000	0.5625	4.2222	0.0000	0.2500	0.0000	0.0000
53	0.5278	4.1494	0.0000	2.2222	2.5347	0.0000	2.5000	0.0000	0.0000	0.0000
54	0.1181	4.0906	0.5625	0.0000	3.3472	3.3611	0.0000	0.0000	0.0000	0.0000
55	0.1078	4.1772	0.4722	0.0000	3.2569	3.5000	0.0000	0.0000	0.0000	0.0000
56	1.2658	2.1319	1.8611	3.3125	1.0000	1.2500	0.3611	0.0000	1.0000	0.0000

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
57	1.1686	2.5486	1.5833	3.1736	0.2500	2.0000	1.1111	0.0000	0.2500	0.0000
58	1.2400	1.8056	2.1875	3.1250	1.0000	1.2500	1.2500	0.0000	0.1111	0.0000
59	1.1494	2.3194	2.5625	2.6111	1.0000	0.3611	1.2500	0.0000	1.0000	0.0000
60	0.9619	1.4861	6.0069	0.0000	0.0000	1.3611	0.0000	2.2500	0.0000	0.0000
61	1.0072	0.9722	6.3333	0.0000	0.0000	2.2500	0.0000	1.3611	0.0000	0.0000
62	1.0522	2.7361	2.2847	2.4722	0.2500	1.1111	2.0000	0.0000	0.2500	0.0000
63	1.0331	3.2083	1.8611	2.6111	0.1111	1.2500	1.2500	0.0000	1.0000	0.0000
64	0.4950	3.0556	3.0347	0.0000	1.3611	2.8611	0.0000	1.0000	0.0000	0.0000
65	0.4725	3.1944	2.8472	0.0000	0.7500	4.2222	0.0000	0.2500	0.0000	0.0000
66	0.4272	4.8056	0.0000	1.6111	1.9722	0.0000	3.2500	0.0000	0.0000	0.0000
67	0.3786	3.9444	2.3333	0.0000	0.4722	3.7500	0.0000	1.0000	0.0000	0.0000
68	2.0764	1.8333	0.0000	6.6111	0.0000	0.0000	1.2500	0.0000	0.0000	1.0000
69	1.9306	0.0000	4.8333	3.4722	0.0000	0.0000	0.0000	1.0000	1.2500	0.0000
70	2.0625	1.5000	0.0000	6.6667	0.0000	0.0000	2.0000	0.0000	0.0000	0.2500

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
71	1.8889	0.0000	5.4444	3.0000	0.0000	0.0000	0.0000	0.2500	2.0000	0.0000
72	1.1875	2.8889	1.2500	3.2222	0.2500	2.0000	0.5000	0.0000	1.0000	0.0000
73	1.0069	3.1667	2.4722	2.0000	0.2500	0.5000	2.0000	0.0000	1.0000	0.0000
74	1.0000	1.4444	5.9167	0.0000	0.0000	1.5000	0.0000	2.2500	0.0000	0.0000
75	0.3750	5.3333	0.0000	1.0000	1.5000	0.0000	4.0000	0.0000	0.0000	0.0000
76	0.0123	3.0548	0.0000	0.0000	9.5015	0.0000	0.0000	0.0000	0.0000	0.0000
77	0.2813	2.5150	2.0204	0.0000	6.4983	1.4914	0.0000	0.0000	0.0000	0.0000
78	0.1471	3.4872	1.2778	0.0000	5.5261	2.4636	0.0000	0.0000	0.0000	0.0000
79	0.1059	3.6972	1.1511	0.0000	5.3161	2.6736	0.0000	0.0000	0.0000	0.0000
80	0.0956	3.7458	1.1250	0.0000	5.2675	2.7222	0.0000	0.0000	0.0000	0.0000
81	0.8112	2.9378	0.0000	3.0278	5.0347	0.0000	1.4636	0.0000	0.0000	0.0000
82	0.5408	2.3700	3.8178	0.0000	3.6111	1.8872	0.0000	1.0000	0.0000	0.0000
83	0.4558	2.8353	3.4236	0.0000	2.6736	3.5747	0.0000	0.2500	0.0000	0.0000
84	0.4333	2.9325	3.3264	0.0000	2.5347	3.8525	0.0000	0.1111	0.0000	0.0000

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
85	0.4575	2.7900	3.3978	0.0000	2.6736	3.7622	0.0000	0.0625	0.0000	0.0000
86	0.5816	1.8544	4.1111	0.0000	3.6111	2.8472	0.0000	0.0400	0.0000	0.0000
87	0.5482	4.1200	0.0000	2.2900	4.0747	0.0000	2.4236	0.0000	0.0000	0.0000
88	0.3729	3.5175	2.9636	0.0000	2.4636	3.0347	0.0000	1.0000	0.0000	0.0000
89	0.3243	3.7953	2.6858	0.0000	1.7136	4.5347	0.0000	0.2500	0.0000	0.0000
90	0.3382	3.7050	2.7050	0.0000	1.7622	4.6250	0.0000	0.1111	0.0000	0.0000
91	0.4754	4.3786	0.0000	2.1736	3.8872	0.0000	2.6111	0.0000	0.0000	0.0000
92	0.3365	3.7050	2.8472	0.0000	2.2761	3.2222	0.0000	1.0000	0.0000	0.0000
93	0.1033	4.1842	0.5400	0.0000	4.7883	3.4236	0.0000	0.0000	0.0000	0.0000
94	0.0882	4.2989	0.4236	0.0000	4.6719	3.6111	0.0000	0.0000	0.0000	0.0000
95	1.2067	2.4256	1.8125	3.2900	2.2500	1.3611	0.4236	0.0000	1.0000	0.0000
96	1.0833	2.9811	1.4444	3.1511	1.3611	2.2500	1.1736	0.0000	0.2500	0.0000
97	1.0644	2.9881	1.4375	3.1025	1.3611	2.2500	1.3125	0.0000	0.1111	0.0000
98	1.1667	2.1181	2.1200	3.0800	2.2500	1.3611	1.3611	0.0000	0.0625	0.0000

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
99	1.0800	2.6356	2.5400	2.5625	2.2500	0.4236	1.3611	0.0000	1.0000	0.0000
100	0.8664	1.9897	5.7969	0.0000	1.0000	1.7847	0.0000	2.2500	0.0000	0.0000
101	0.8214	2.2258	5.5122	0.0000	0.2500	3.4236	0.0000	1.3611	0.0000	0.0000
102	0.9133	1.4306	6.0975	0.0000	1.0000	2.8611	0.0000	1.1736	0.0000	0.0000
103	0.9306	3.2397	2.2622	2.3333	1.3611	1.1736	2.2500	0.0000	0.2500	0.0000
104	0.7953	2.3647	5.4219	0.0000	0.1111	3.5625	0.0000	1.3611	0.0000	0.0000
105	0.8611	1.7500	5.8267	0.0000	0.2500	4.2222	0.0000	0.5625	0.0000	0.0000
106	0.9378	3.1981	2.1650	2.3750	1.3611	1.3125	2.2500	0.0000	0.1111	0.0000
107	0.9272	3.7119	1.7222	2.5625	1.1736	1.5000	1.3611	0.0000	1.0000	0.0000
108	0.8561	3.9897	1.5347	2.4236	0.5625	2.1111	2.1111	0.0000	0.2500	0.0000
109	0.9011	3.7606	1.8125	2.4722	1.1736	1.3611	1.5000	0.0000	1.0000	0.0000
110	0.7397	2.9272	5.0694	0.0000	0.0625	2.7222	0.0000	2.2500	0.0000	0.0000
111	0.4583	3.2642	2.9636	0.0000	2.6736	3.0347	0.0000	1.0000	0.0000	0.0000
112	0.4256	3.4897	2.6858	0.0000	1.9722	4.5347	0.0000	0.2500	0.0000	0.0000

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
113	0.4236	3.4517	2.7050	0.0000	1.9722	4.6250	0.0000	0.1111	0.0000	0.0000
114	0.3803	5.1617	0.0000	1.5625	3.3472	0.0000	3.3611	0.0000	0.0000	0.0000
115	0.3214	4.2883	2.1458	0.0000	1.6458	4.1111	0.0000	1.0000	0.0000	0.0000
116	0.3092	4.3406	2.0486	0.0000	1.1250	5.3333	0.0000	0.2500	0.0000	0.0000
117	0.3056	4.3406	2.1458	0.0000	1.5972	4.1111	0.0000	1.0000	0.0000	0.0000
118	0.3439	5.1494	0.0000	1.4722	3.2569	0.0000	3.5000	0.0000	0.0000	0.0000
119	0.0833	4.3650	0.3333	0.0000	4.6042	3.7500	0.0000	0.0000	0.0000	0.0000
120	0.4378	3.3994	2.8472	0.0000	2.5347	3.2222	0.0000	1.0000	0.0000	0.0000
121	1.9117	2.4931	0.0000	6.4236	1.0000	0.0000	1.6111	0.0000	0.0000	1.0000
122	1.6756	1.5486	3.9444	3.4236	0.0000	1.2500	0.1111	1.0000	1.2500	0.0000
123	1.6983	1.0833	4.2708	3.3750	0.0000	2.0000	0.2500	0.2500	1.1111	0.0000
124	1.7172	3.0486	0.0000	5.8681	0.2500	0.0000	3.1111	0.0000	0.0000	0.2500
125	1.6497	1.2222	4.2708	3.2361	0.0000	1.2500	1.0000	1.0000	0.3611	0.0000
126	1.8600	2.1667	0.0000	6.3750	1.0000	0.0000	2.5000	0.0000	0.0000	0.1111

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
127	1.6078	1.5972	4.6458	2.8611	0.0000	1.1111	0.2500	0.2500	2.0000	0.0000
128	1.6044	1.2222	4.9722	2.7222	0.0000	1.2500	1.0000	0.1111	1.2500	0.0000
129	1.5592	1.7361	4.6458	2.7222	0.0000	0.3611	1.0000	1.0000	1.2500	0.0000
130	1.4656	0.0000	8.9167	0.0000	0.0000	0.0000	0.0000	3.6111	0.0000	0.0000
131	1.6789	3.5694	0.0000	5.7222	0.1111	0.0000	2.5000	0.0000	0.0000	1.0000
132	1.0975	3.3403	1.1111	3.1736	1.3611	2.2500	0.6111	0.0000	1.0000	0.0000
133	1.0525	3.5694	0.8333	3.1250	0.7500	3.0000	1.2222	0.0000	0.2500	0.0000
134	0.8908	3.6667	2.4236	1.8611	1.3611	0.6111	2.2500	0.0000	1.0000	0.0000
135	0.8197	2.6944	5.0694	0.0000	0.2500	2.7222	0.0000	2.2500	0.0000	0.0000
136	0.7936	2.8333	4.9792	0.0000	0.1111	2.8611	0.0000	2.2500	0.0000	0.0000
137	0.8197	3.9444	2.2361	1.7222	0.7500	1.2222	3.0000	0.0000	0.2500	0.0000
138	0.8650	2.1806	5.3958	0.0000	0.2500	3.6111	0.0000	1.3611	0.0000	0.0000
139	0.8647	4.4167	1.1111	2.4722	0.4722	2.2500	1.5000	0.0000	1.0000	0.0000
140	0.7744	4.5556	1.7222	1.8611	0.4722	1.5000	2.2500	0.0000	1.0000	0.0000

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
141	0.3075	5.6250	0.0000	0.8611	2.8333	0.0000	4.2500	0.0000	0.0000	0.0000
142	0.2850	4.9444	1.4444	0.0000	0.9444	5.0000	0.0000	1.0000	0.0000	0.0000
143	0.9100	1.9444	5.6806	0.0000	1.0000	1.9722	0.0000	2.2500	0.0000	0.0000
144	2.7361	0.0000	2.8333	6.9722	0.0000	0.0000	0.0000	0.0000	1.2500	1.0000
145	2.7292	0.0000	2.5000	7.1667	0.0000	0.0000	0.0000	0.0000	2.0000	0.2500
146	1.7431	3.3889	0.0000	5.7222	0.2500	0.0000	2.5000	0.0000	0.0000	1.0000
147	1.6875	1.5556	3.9444	3.3333	0.0000	1.2500	0.2500	1.0000	1.2500	0.0000
148	1.5556	1.6944	5.1667	2.2500	0.0000	0.5000	1.0000	0.2500	2.0000	0.0000
149	0.8125	4.3333	2.3333	1.2500	0.7500	0.7500	3.0000	0.0000	1.0000	0.0000
150	1.5000	0.0000	8.8333	0.0000	0.0000	0.0000	0.0000	3.7500	0.0000	0.0000

表 4-2 MEDV 与六种其他拓扑指数对于烷烃异构体的分辨率*

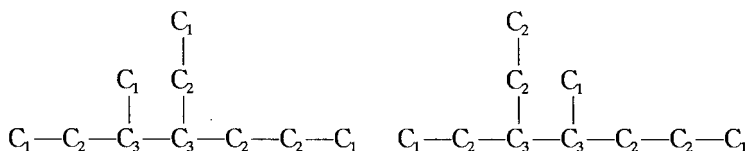
描述子	庚烷 ($N_T = 9$)		辛烷 ($N_T = 18$)		壬烷 ($N_T = 35$)		癸烷 ($N_T = 75$)	
	N_R	R_D	N_R	R_D	N_R	R_D	N_R	R_D
Wiener (W)	7	0.78	16	0.89	16	0.46	40	0.53
Hosoya (Z)	9	1.00	14	0.78	23	0.66	40	0.53
Randic ($^1\chi$)	9	1.00	16	0.89	28	0.80	49	0.65
Bonchev (I_D)	9	1.00	18	1.00	33	0.94	71	0.95
Balaban (J)	9	1.00	18	1.00	35	1.00	74	0.99
Gutman (M_1)	5	0.56	6	0.33	7	0.20	8	0.11
MEDV - 4	9	1.00	18	1.00	35	1.00	75	1.00

N_R 是指能分辨的异构体数, N_T 为总的异构体数目, 此时 $R_D = N_R/N_T$ 。

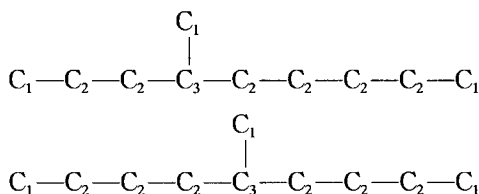
正如第三章所论述的, 以各样本分子在 MEDV - 4 矢量张成的 10 维模式空间中的模式距离作为量度, 可方便地进行分子相似性分析。两样本分子的模式距离越小, 表明两分子的相似性越大。将表 4-1B 中烷烃的各个 MEDV - 4 分子描述子数值代入方程 (3-1) 式, 可以计算各对烷烃分子之间的模式距离。计算结果表明 模式距离在 0.50 以内的分子只占 0.4% (也可参见表 3-1), 即只有 44 对烷烃分子 (总计 11 175 对) 间的模式距离小于或等于 0.50。同样, 模式距离小于 0.25 的分子对, 有 21 对, 其中小于 0.20 的只有 9 对。这些距离 D (分子 1, 分子 2) 是:

$$\begin{aligned}
 D(115, 117) &= 0.0731, & D(79, 80) &= 0.0887, \\
 D(130, 150) &= 0.1656, & D(122, 147) &= 0.1662, \\
 D(112, 113) &= 0.1711, & D(60, 74) &= 0.1750, \\
 D(94, 119) &= 0.1909, & D(114, 118) &= 0.1926, \\
 D(89, 90) &= 0.1963
 \end{aligned}$$

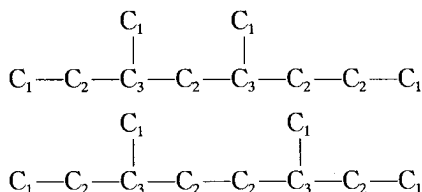
其中距离最小的是表 4-1 中 115 号与 117 号烷烃分子之间的模式距离: $D(115, 117) = 0.0731$ 。这表明 115 号分子的 3-甲基-4-乙基庚烷和 117 号分子的 4-甲基-3-乙基庚烷之间具有的高度相似性:



显然它们的分子结构是非常相似的。模式距离稍大的 ($D = 0.0887$) 是 79 号的 4-甲基壬烷和 80 号的 5-甲基壬烷 其分子结构的相似程度也是很大的。



模式距离再大些的 ($D = 0.1963$) 要数 89 号的 3,5-二甲基辛烷和 90 号的 3,6-二甲基辛烷, 其分子结构差异也要大些:



4.1.2 150 个烷烃沸点 QSPR 模型

物质的沸点是定量结构-性质相关 (QSPR) 研究中最常用到的性质之一^[7~12]。本节较详细地讨论 MEDV-4 在烷烃沸点 QSPR 研究中的应用。

150 个烷烃包括由 1 个碳直到 10 个碳原子构成的全部烷烃异构体。它们的沸点实验观测值取自文献^[13] 其范围从 -164 到 174.12°C , 其实验误差范围一般在 $5.4 \sim 11.4\text{ K}$ 之间^[11] 各个烷烃的实验沸点列入表 4-3 中。150 个烷烃的 MEDV-4 矢量分子描述子参见表 4-1B^[5]。

表 4-3 150 个烷烃的实验沸点 $b. p.$ (单位: $^{\circ}\text{C}$)

No	$b. p.$	No	$b. p.$	No	$b. p.$	No	$b. p.$	No	$b. p.$
1	-164.00	19	80.500	37	113.400	55	141.200	73	141.600
2	-88.600	20	86.100	38	115.600	56	131.700	74	136.730
3	-42.100	21	93.500	39	118.200	57	126.500	75	146.200
4	-0.500	22	80.900	40	106.500	58	124.000	76	174.120
5	-11.700	23	125.700	41	150.770	59	137.700	77	167.000
6	36.100	24	117.600	42	142.800	60	139.000	78	167.800
7	27.800	25	118.000	43	143.800	61	131.300	79	165.700
8	9.500	26	117.700	44	142.400	62	126.500	80	165.100
9	69.000	27	106.800	45	132.700	63	140.500	81	155.000
10	60.300	28	115.600	46	140.500	64	138.000	82	164.310
11	63.300	29	109.400	47	133.500	65	133.800	83	153.000
12	49.700	30	109.000	48	136.000	66	140.600	84	156.800
13	58.000	31	112.000	49	135.200	67	140.400	85	158.540
14	98.400	32	117.700	50	137.300	68	140.270	86	159.870
15	90.000	33	118.500	51	140.100	69	133.000	87	161.200
16	92.000	34	110.000	52	136.000	70	122.700	88	166.000
17	79.200	35	99.200	53	135.200	71	141.500	89	160.000
18	89.800	36	114.700	54	143.000	72	133.830	90	160.000

续表

No	b. p.	No	b. p.	No	b. p.	No	b. p.	No	b. p.
91	157.500	103	153.000	115	167.000	127	164.590	139	165.000
92	162.100	104	157.000	116	158.300	128	153.000	140	170.000
93	166.000	105	144.800	117	167.000	129	162.200	141	166.300
94	163.640	106	152.800	118	167.000	130	161.000	142	162.000
95	158.000	107	164.000	119	160.000	131	170.500	143	163.000
96	147.700	108	155.680	120	160.000	132	159.000	144	166.050
97	148.000	109	164.000	121	158.000	133	147.000	145	159.290
98	148.200	110	164.000	122	154.900	134	169.000	146	168.000
99	160.100	111	166.000	123	148.400	135	164.000	147	155.300
100	163.000	112	160.000	124	153.300	136	164.000	148	169.440
101	157.000	113	159.700	125	148.200	137	158.000	149	174.000
102	155.700	114	163.800	126	137.460	138	157.000	150	157.040

以多元线性回归 MLR 程序计算 MEDV 描述子与 150 个烷烃沸点之间的相关模型，并计算有关统计参量。为了统一，先将摄氏沸点()加上 273.15 变换为开氏沸点 (K)，然后建立沸点 ($b.p.$) 与 MEDV-4 描述子 (x_i) 之间的 QSPR 模型^[5]：

$$\begin{aligned} b.p. = & (107.0969 \pm 4.9478) + (92.0674 \pm 6.3747)x_1 + \\ & (62.2532 \pm 1.9238)x_2 + (30.5024 \pm 0.8980)x_3 + \\ & (7.8862 \pm 1.4796)x_4 + (15.5087 \pm 0.3766)x_5 + \\ & (8.9157 \pm 1.1143)x_6 - (23.8692 \pm 1.7506)x_7 - \\ & (21.5509 \pm 1.8932)x_8 - (28.0297 \pm 2.3907)x_9 - \\ & (25.1267 \pm 3.8830)x_{10} \end{aligned} \quad (4-1)$$

$$n = 150, m = 10, r = 0.9951, RMSE = 4.815 \text{ K}, F = 1407.835$$

(建模时)

$$q = 0.9829, RMSV = 9.924 \text{ K} \quad (\text{LOO 校验})$$

式中 $n, m, r, RMSE, F$ 分别是建模过程中的样本数、变量数、相关系数、均方根误差以及 Fischer 统计量 F 。括号内“ $\pm$ ”号后面的数据为各回归系数的标准偏差。由 $r = 0.9951$ 以及 $RMSE = 4.815 \text{ K}$ 可知 150 个烷烃化合物的沸点与 10 个 MEDV-4 描述子之间具有良好的相关性，表明模型 (4-1) 式具有对内部样本的良好估计能力。

为了说明模型稳定性和模型对外部样本分子理化性质（如沸点）的预测能力，应进行交互检验。LOO (leave-one-out) 交互检验法是最常使用和最有效的说明模型稳定性的方法之一。在样本数相当多的情况下（比如样本数远大于模型中所含变量数的 6 倍以上时），LOO 结果统计量也是说明模型预测能力的重要依据之一。(4-1) 式中，LOO 统计量（计算方法参见第三章）以 LOO 相关系数 (q) 和 LOO 检验均方根误差 ($RMSV$) 等表示。对 150 个烷烃沸点的 LOO 交互检验结果 $q = 0.9829, RMSV = 9.924 \text{ K}$ 表明，除开始的第 1, 2, 3 号烷烃的 LOO 预测值明显偏大以及第 4 号偏

大之外，其他都较小。1号甲烷预测误差达 -96.38 K 特别大，这是由于甲烷分子结构中没有碳原子类型，所有 MEDV-4 描述子值均为零，区别于其他任何烷烃分子；2号乙烷只有第 1 类碳原子，MEDV-4 中只有第 1 个描述子 x_1 不等于零，误差为 39.75 K ，为次大；3号丙烷和 4号丁烷只有第 1 和第 2 原子类型，MEDV-4 矢量中只有描述子 x_1 和 x_2 不等于零，误差分别为 26.41 K 与 17.67 K ，较大。若去除这四个离群值，重建沸点与 MEDV-4 分子描述子的 QSPR 模型，得：

$$\begin{aligned}
 b.p. = & (62.4382 \pm 24.1899) + (171.3715 \pm 24.7920)x_1 + \\
 & (78.9076 \pm 8.5458)x_2 + (28.0933 \pm 1.8148)x_3 - \\
 & (8.7966 \pm 3.6552)x_4 + (14.9306 \pm 0.3458)x_5 + \\
 & (16.7059 \pm 4.0600)x_6 - (36.0998 \pm 6.6261)x_7 - \\
 & (35.4258 \pm 7.1106)x_8 - (46.3556 \pm 9.3519)x_9 - \\
 & (49.7318 \pm 11.8764)x_{10} \quad (4-2)
 \end{aligned}$$

$$n = 146, m = 10, r = 0.9953, \text{RMSE} = 3.282\text{ K}, F = 1426.961$$

(建模时)

$$q = 0.9938, \text{RMSV} = 3.758\text{ K} \quad (\text{LOO 校验})$$

上式结果表明，1~4 号样本的加入对模型影响很大，去除后 LOO 均方根误差 3.758 K 明显降低，说明模型稳定性明显变好。

图 4-1 表明了实验观测沸点 (BP_{OBS}) 与由 (4-2) 式估计的沸点 (BP_{EST}) 及 LOO 交互检验预测的沸点 (BP_{LOO}) 之间的总体相关结果。

此外，我们注意到模型 (4-1) 式中第 10 和第 4 个描述子 (x_{10}, x_4) 回归系数的标准偏差较大，显著性欠佳。分析 MEDV-4 中各个描述子可知，第 10 个描述子在 150 个样本中只有 10 个烷烃分子不为零，而第 4 个描述子虽有 67 个非零值，但都与第四原子类型有关，可以认为造成这两个回归系数统计上不很显著的原因是 150 个烷烃中属于第四原子类型的原子数较少。所以，在

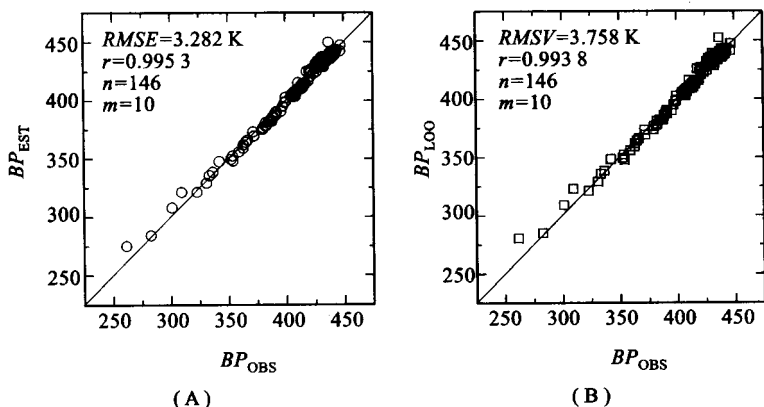


图 4-1 模型估计 (A)与模型 LOO 校验(B) 沸点与实验沸点相关图

对碳原子数更大烷烃的沸点进行估计与预测时，还应提高这两个回归系数的统计显著性，在建立模型时应切实注意引入更多的含第四类碳原子的烷烃分子作为训练集样本，以期更准确地估计与预测更多烷烃化合物的沸点。应用最佳子集回归 BSR 程序进行变量优化选择时，也发现 x_{10}, x_4 的显著性低，不包括这两个描述子的模型稳定性更好，其 LOO 相关系数 q 更高，LOO 均方根误差 RMSV 更小。有关统计量分别为：

$$n = 150, m = 8, r = 0.9922, RMSE = 6.083 \text{ K}, F = 1111.825 \quad (\text{建模时})$$

$$q = 0.9902, RMSV = 6.816 \text{ K} \quad (\text{LOO 校验})$$

$$n = 146, m = 8, r = 0.9944, RMSE = 3.595 \text{ K}, F = 1506.182 \quad (\text{建模时})$$

$$q = 0.9931, RMSV = 3.970 \text{ K} \quad (\text{LOO 校验})$$

4.1.3 沸点模型的比较分析

上节我们建立了 150 个烷烃沸点与 MEDV-4 矢量描述子的

定量相关模型，并对模型估计能力与预测能力（稳定性）进行了评价，证明了 MEDV-4 适合于烷烃分子沸点的估计与预测。本节进一步将 MEDV-4 方法与其他方法作比较分析。

Katritzky 等在最近一篇评述有关结构多样性分子理化性质的 QSPR 研究工作中，对有关化合物的沸点估计与预测进行了比较全面的评述^[14]。目前之最好结果当推 Wessel 和 Jurs 在 1995 年所发表的工作^[12]，即使用神经网络技术对 269 个烃类化合物的沸点与 6 个曾用于 MLR 方法的重要描述子之间进行模拟研究，在网络节点数为 6:5:1 的三层网络结构条件下得到训练集建模时的 RMS 值为 5.7 K（本文:RMSE = 4.815 K）这个结果稍稍优于 MLR 结果（6.3 K）。应当指出，由于相关系数与模型所含描述子个数及使用的样本数目有关且目前尚无明显的规律，因而很难通过其大小对不同方法、不同训练集样本数以及不同描述子所建立的各种模型进行直接比较以说明方法的好坏。同样平均相对误差与沸点所用单位（K 或 °C）有关，且不能反映误差大样本的严重影响，作为模型比较似有点勉强。另一方面，正如我们在第三章模型质量评价中所讨论的，一个模型的质量好坏不仅在于模型估计误差的大小，而且更重要的在于模型对外部样本性质的预测能力大小。然而，由于不同文献对说明预测能力方面所用方法不相一致，而且采用什么方法最为合理，还没有统一的理论说明，有些文献根本就没有这方面的数据。所以，目前尚难做出非常可靠的比较分析，只能在模型估计能力方面做些讨论。

由于烷烃沸点的大小随碳链的增长，沸点增加的趋势越来越平缓，以沸点对分子中碳原子数作的散点图形似对数曲线。因此，烷烃沸点经对数变换后再与 MEDV-4 之间建立的线性关系模型可能更好。结合作者先前研究 λ 矢量（MEDV 矢量的最初形式）方法的结果^[15]，重建对数模型如下：

$$\ln(539.85 - b \cdot p) = (6.5589 \pm 0.0077) - (0.1328 \pm 0.0099)x_1$$

$$\begin{aligned}
& (0.0939 \pm 0.0030)x_2 - (0.0510 \pm 0.0014)x_3 - \\
& (0.0181 \pm 0.0023)x_4 - (0.0382 \pm 0.0006)x_5 - \\
& (0.0044 \pm 0.0017)x_6 + (0.0135 \pm 0.0027)x_7 + \\
& (0.0071 \pm 0.0029)x_8 - (0.0079 \pm 0.0037)x_9 - \\
& (0.0124 \pm 0.0060)x_{10} \quad (4-3)
\end{aligned}$$

$$n = 150, m = 10, r = 0.9965, RMSE = 3.22 \text{ K}, F = 1949.31$$

(建模时)

$$q = 0.9888, RMSV = 12.38 \text{ K} \quad (\text{LOO 校验})$$

显然, 相关系数 r 提高 均方根误差 $RMSE$ 明显降低 统计显著性增强 (F 变大)。经 LOO 交互检验法计算所得的 $RMSV = 12.38 \text{ K}$ 虽然 $RMSV$ 较方程 (4-1) 还大, 但这主要是由于第 1 个样本甲烷的误差 (145.01 K) 太大的缘故。即使这样, 这个模型也已具有相当好的稳定性与预测能力了 (一般情况下, 预测能力还应该以多个检验集进行检验, 这里由于样本数远大于模型变量数的 5 倍, LOO 结果已能大致说明预测能力的好坏)。这与文献 λ 矢量结果^[15] ($r = 0.9948$ 和 $RMSV = 4.985 \text{ K}$) 相比, 也有明显改善。

若去除前面三个误差明显偏大的甲烷、乙烷与丙烷样本, 则得到稳定性与预测能力大大提高的 QSPR 相关模型:

$$\begin{aligned}
\ln(539.85 - b.p.) = & (5.6588 \pm 0.1674) + (0.1936 \pm 0.1717)x_1 + \\
& (0.0121 \pm 0.0593)x_2 - (0.0551 \pm 0.0126)x_3 - \\
& (0.0880 \pm 0.0253)x_4 - (0.1190 \pm 0.0024)x_5 - \\
& (0.1470 \pm 0.0282)x_6 + (0.1875 \pm 0.0460)x_7 + \\
& (0.2158 \pm 0.0493)x_8 - (0.2934 \pm 0.0648)x_9 - \\
& (0.4346 \pm 0.0823)x_{10} \quad (4-4)
\end{aligned}$$

$$n = 147, m = 10, r = 0.9952, RMSE = 2.90 \text{ K}, F = 1418.39$$

(建模时)

$$q = 0.9941, RMSV = 3.34 \text{ K} \quad (\text{LOO 校验})$$

由以上分析结果可知，MEDV 方法的建模与预测误差都很小，表明模型具有良好估计与预测能力。可用于烷烃化合物沸点的估计与预测。

4.2 烷烃其他性质 QSPR 模型

以与研究沸点一样的方法研究烷烃密度、折光指数、临界温度、临界压力、蒸发热、热容、熵等 10 余种物理化学性质、热力学函数以及色谱保留指数与 MEDV-4 矢量描述子的定量相关关系，结果分述如下。

4.2.1 密度与折光指数

1 atm ($1 \text{ atm} = 1.01 \times 10^5 \text{ Pa}$) 下 25°C 时的密度 density (简称 *Den*) 数据是表征与识别化合物的重要物理化学性质之一。同时密度还可以用于预测与估计化合物的其他物理化学性质比如临界压力等。折光指数 (refractive index, 简称 *RI*) 或折光率定义为光通过真空的速度与光通过所研究物质的速度比值，是最重要的光学性质之一，常常用来表征有机化合物。因而建立与 MEDV-4 描述子的相关模型不仅可以说明 MEDV-4 描述子与有关理化性质的相关关系，而且也具有实际应用价值。

对于密度与折光指数，选用相同的数据集。选择 Gakh 报告^[16] 中的 134 个烷烃及相应密度与折光指数，应用 MLR 方法建立密度及折光指数与 MEDV-4 的相关模型以 LOO 交互检验证明模型稳定性及预测能力。134 个烷烃的密度与折光指数实验值参见表 4-4 中。

由于第 10 个 MEDV 描述子在 134 个烷烃中只有 10 个不为零，仅占 7.46%，因此删去该描述子，应用 MLR 技术建立 9 个 MEDV 描述子的密度模型：

表 4-4 134 个烷烃的实验密度 Den (单位: g/cm^3) 与折光指数 RI 值

No	Den	RI	No	Den	RI	No	Den	RI	No	Den	RI
10	648.52	1.3687	28	708.16	1.3988	47	711.50	1.4011	65	724.20	1.4054
11	659.76	1.3739	29	696.17	1.3929	48	713.60	1.4015	66	736.00	1.4134
12	644.46	1.3660	30	689.37	1.3900	49	704.50	1.3985	67	738.00	1.4128
13	657.02	1.3723	31	705.95	1.3978	50	721.60	1.4063	68	752.97	1.4214
14	679.50	1.3851	32	715.15	1.4018	51	727.50	1.4091	69	735.22	1.4125
15	674.34	1.3823	33	709.45	1.3992	52	716.60	1.4046	70	715.61	1.4046
16	682.88	1.3861	34	712.03	1.4007	53	718.30	1.4053	71	751.11	1.4200
17	669.48	1.3800	35	687.84	1.3890	54	722.50	1.4070	72	731.00	1.4101
18	690.81	1.3895	36	722.30	1.4052	55	722.30	1.4067	73	750.80	1.4197
19	668.23	1.3788	37	715.09	1.4020	56	725.70	1.4082	74	734.10	1.4115
20	689.16	1.3884	38	715.20	1.4017	57	711.80	1.4010	75	749.92	1.4184
21	693.92	1.3971	39	723.54	1.4055	58	703.20	1.3973	81	720.80	1.4060
22	685.64	1.3869	40	721.70	1.4057	59	733.50	1.4119	82	734.40	1.4127
23	698.54	1.3951	42	709.60	1.4008	60	735.10	1.4120	83	722.60	1.4069
24	693.87	1.3926	43	716.70	1.4040	61	717.90	1.4037	84	726.40	1.4089
25	701.73	1.3961	44	716.30	1.4041	62	720.05	1.4052	85	723.60	1.4084
26	700.71	1.3955	45	706.60	1.3995	63	741.40	1.4154	86	720.20	1.4062
27	691.11	1.3910	46	722.00	1.4064	64	729.00	1.4091	87	735.10	1.4142

续表

No	Den	RI	No	Den	RI	No	Den	RI	No	Den	RI
88	741.80	1.4159	104	737.30	1.4160	120	735.40	1.4132	137	751.40	1.4202
89	732.90	1.4115	105	719.00	1.4071	121	760.89	1.4260	138	736.80	1.4232
90	732.90	1.4115	106	736.20	1.4149	122	751.30	1.4193	139	759.80	1.4246
91	731.20	1.4122	107	752.70	1.4236	123	733.60	1.4119	140	759.60	1.4244
92	743.20	1.4167	108	739.00	1.4170	124	742.40	1.4185	141	757.50	1.4235
93	735.90	1.4136	109	753.50	1.4235	125	731.60	1.4132	142	747.20	1.4167
94	734.30	1.4131	110	751.90	1.4229	126	714.80	1.4055	143	743.60	1.4172
95	738.50	1.4145	111	739.80	1.4151	127	765.60	1.4298	144	776.75	1.4341
96	725.70	1.4092	112	732.20	1.4114	128	744.90	1.4196	145	763.61	1.4281
97	728.10	1.4101	113	731.80	1.4111	129	758.60	1.4267	146	778.00	1.4397
98	723.80	1.4178	114	746.30	1.4185	130	745.60	1.4204	147	753.10	1.4199
99	748.80	1.4202	115	746.60	1.4183	131	778.90	1.4368	148	773.50	1.4310
100	748.50	1.4195	116	736.80	1.4141	132	744.70	1.4174	149	775.50	1.4320
101	745.10	1.4169	117	746.80	1.4184	133	730.20	1.4107	150	754.57	1.4246
102	734.70	1.4131	118	747.20	1.4187	134	759.90	1.4247			
103	734.60	1.4143	119	732.10	1.4113	135	751.60	1.4203			

$$\begin{aligned}
Den = & (466.4585 \pm 5.7203) + (150.5954 \pm 7.6421)x_1 + \\
& (58.8993 \pm 1.9905)x_2 + (14.4457 \pm 1.2730)x_3 - \\
& (16.9197 \pm 2.0929)x_4 + (8.2611 \pm 0.4147)x_5 - \\
& (11.9181 \pm 1.0948)x_6 - (21.6019 \pm 1.6416)x_7 - \\
& (17.0126 \pm 1.8528)x_8 - (15.2123 \pm 1.9206)x_9 \quad (4-5)
\end{aligned}$$

$n = 134, m = 9, r = 0.9882, RMSE = 4.016 \text{ g/cm}^3, F = 571.496$
(建模时)

$$q = 0.9856, RMSV = 4.423 \text{ g/cm}^3 \quad (\text{LOO 校验})$$

结果表明, 模型具有良好估计与预测能力。如同烷烃沸点, 密度也随烷烃碳原子数增加, 变得越来越缓慢。因此, 使用对数变换应该能提高模型的质量。重建 9 个 MEDV-4 描述子与密度的对数模型如下:

$$\begin{aligned}
\ln(867 - Den) = & (6.8278 \pm 0.0308) - (1.2096 \pm 0.0411)x_1 - \\
& (0.4576 \pm 0.0107)x_2 - (0.1022 \pm 0.0069)x_3 + \\
& (0.1550 \pm 0.0113)x_4 - (0.0494 \pm 0.0022)x_5 + \\
& (0.1139 \pm 0.0059)x_6 + (0.1882 \pm 0.0460)x_7 + \\
& (0.1639 \pm 0.0100)x_8 + (0.1419 \pm 0.0103)x_9 \quad (4-6)
\end{aligned}$$

$n = 134, m = 9, r = 0.9930, RMSE = 2.876 \text{ g/cm}^3, F = 970.326$
(建模时)

$$q = 0.9914, RMSV = 3.150 \text{ g/cm}^3 \quad (\text{LOO 校验})$$

上述结果与 (4-5) 式模型相比, 相关系数 r 从 0.9882 上升到 0.9930, 均方根误差 $RMSE$ 从 4.016 下降到 2.876 g/cm^3 , F 从 571.496 上升到 970.326。LOO 交互检验的有关统计结果为: $q = 0.9914$ 以及 $RMSV = 3.150 \text{ g/cm}^3$ 。比原来 LOO 结果 $q = 0.9856$ 与 $RMSV = 4.423 \text{ g/cm}^3$ 有较大改善。

对折光指数进行类似研究, 建立的 9 参数对数模型如下:

$$\begin{aligned}
\ln(1.4918 - RI) = & -(0.8527 \pm 0.0380) - (0.9575 \pm 0.0507)x_1 - \\
& (0.3881 \pm 0.0132)x_2 - (0.1041 \pm 0.0084)x_3 +
\end{aligned}$$

$$(0.102\ 8 \pm 0.013\ 9)x_4 - (0.046\ 7 \pm 0.002\ 8)x_5 + \\ (0.093\ 5 \pm 0.007\ 3)x_6 + (0.157\ 7 \pm 0.010\ 9)x_7 + \\ (0.141\ 3 \pm 0.012\ 3)x_8 + (0.135\ 4 \pm 0.012\ 7)x_9 \\ (4-7)$$

$n = 134, m = 9, r = 0.985\ 8, RMSE = 0.002\ 04, F = 474.578$

(建模时)

$q = 0.982\ 7, RMSV = 0.002\ 24$ (LOO 校验)

上述结果表明 9 参数对数模型具有良好相关性及良好预测能力。

由模型 (4-6) 式估计的密度 (DEN_{EST}) 与实验观测的密度 (DEN_{OBS}) 相关图如图 4-2A 所示。由 (4-7) 式估计的折光指数 (RI_{EST}) 及实验观测折光指数 (RI_{OBS}) 相关图如图 4-2B 所示。

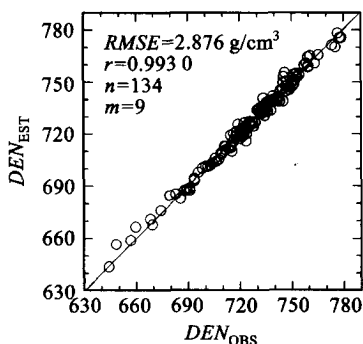


图 4-2A 134 个烷烃估计与实验密度相关图

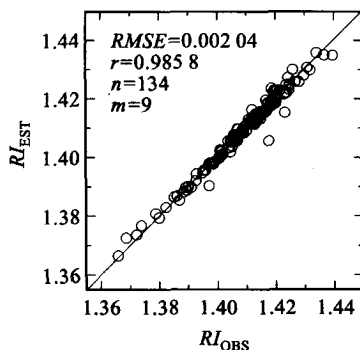


图 4-2B 134 个烷烃估计与实验折光指数相关图

4.2.2 临界温度与临界压力

本节讨论 20 ℃ 时 74 个烷烃的临界温度 (critical temperature,

简称 CT) 和临界压力 (critical pressure 简称 CP) 的 QSPR 模型。烷烃 CT 和 CP 等性质实验观测值直接选自 Needham 等的实验报告^[17], 实验测定的临界温度和临界压力数据参见表 4-5。实验发现, 这两个性质的变化规律如同烷烃沸点、密度与折光指数的变化, 对数变换模型优于常规模型。利用 MLR 方法, 建立对数 QSPR 模型, 并用 LOO 交互检验方法检验其预测能力。结果表明第 9 和第 10 个 MEDV 描述子 (x_9 和 x_{10}) 的回归系数标准偏差偏大, 这是因为 x_{10} 在 74 个烷烃中只有 3 个非零样本, 仅占 4.05%。然而, x_9 有 14 个非零样本, 占 18.92%, 能否删去 x_9 ? 考察比较 MLR 方法建立的 9 参数与 8 参数回归方程可知, 8 参数模型最优, 因而可以删去 x_9 。

对于临界温度 (CT), 其最优 8 参数回归模型为:

$$\begin{aligned} \ln(756.15 - CT) = & (6.635\ 0 \pm 0.030\ 2) - (0.564\ 4 \pm 0.041\ 0)x_1 - \\ & (0.280\ 4 \pm 0.010\ 3)x_2 - (0.113\ 2 \pm 0.004\ 9)x_3 + \\ & (0.012\ 0 \pm 0.010\ 5)x_4 - (0.085\ 1 \pm 0.002\ 9)x_5 + \\ & (0.009\ 3 \pm 0.005\ 8)x_6 + (0.043\ 3 \pm 0.008\ 8)x_7 + \\ & (0.029\ 3 \pm 0.010\ 1)x_8 \end{aligned} \quad (4-8)$$

$$n = 74, m = 8, r = 0.994\ 3, RMSE = 5.30\ K, F = 712.001$$

(建模时)

$$q = 0.988\ 5, RMSV = 10.573\ K \quad (\text{LOO 校验})$$

对于临界压力, 其最优 8 参数回归模型为:

$$\begin{aligned} \ln(CP - 1) = & (3.556\ 5 \pm 0.028\ 5) + (0.291\ 6 \pm 0.038\ 6)x_1 + \\ & (0.025\ 1 \pm 0.009\ 7)x_2 - (0.077\ 4 \pm 0.004\ 6)x_3 - \\ & (0.140\ 4 \pm 0.009\ 9)x_4 - (0.070\ 4 \pm 0.002\ 8)x_5 - \\ & (0.086\ 4 \pm 0.005\ 5)x_6 - (0.068\ 9 \pm 0.008\ 3)x_7 - \\ & (0.052\ 7 \pm 0.009\ 6)x_8 \end{aligned} \quad (4-9)$$

表 4-5 74 个烷烃的临界温度 CT (单位:K)与临界压力 CP (单位:atm, 1 atm = 1.01×10^5 Pa)

No	CT	CP	No	CT	CP	No	CT	CP	No	CT	CP
2	32.27	48.20	21	267.60	28.60	40	270.80	24.50	59	326.10	25.56
3	96.80	42.01	22	258.30	29.75	41	322.00	22.74	60	324.20	25.46
4	152.01	37.47	23	296.20	24.64	42	315.00	23.60	61	309.40	23.49
5	134.98	36.00	24	288.00	24.80	43	318.00	23.70	62	309.10	23.79
6	196.62	33.31	25	292.00	25.60	44	318.30	23.06	63	330.60	26.45
7	187.80	32.90	26	290.00	25.60	45	302.00	22.80	64	322.70	24.77
8	160.60	31.57	27	279.00	25.60	46	315.00	23.79	65	330.30	25.56
9	234.70	29.92	28	293.00	26.60	47	306.00	22.70	66	327.20	25.66
10	224.90	29.95	29	282.00	25.80	48	307.80	22.70	67	312.30	23.59
11	231.20	30.83	30	279.00	25.00	49	306.00	23.70	68	334.50	27.04
12	216.20	30.67	31	290.84	27.20	50	314.00	24.19	69	319.60	25.66
13	227.10	30.99	32	298.00	27.40	51	322.70	24.77	70	301.60	24.58
14	267.01	27.01	33	292.00	25.74	52	312.30	23.59	71	334.50	26.85
15	257.90	27.20	34	294.00	28.20	53	317.80	24.18	72	322.60	25.96
16	262.40	28.10	35	271.15	25.50	54	318.00	23.98	73	338.60	26.94
17	247.70	28.40	36	303.00	29.00	55	318.30	23.98	74	324.20	25.46
18	264.60	29.20	37	295.00	27.60	56	318.10	25.07	75	342.80	26.94
19	247.10	27.40	38	295.00	27.40	57	301.00	23.39			
20	263.00	30.00	39	305.00	28.90	58	296.60	22.41			

$n = 74, m = 8, r = 0.9864, RMSE = 0.591 \text{ atm}, F = 293.640$

(建模时)

$q = 0.9805, RMSV = 0.721 \text{ atm}$ (LOO 校验)

由 (4-8) 式对数模型估计的 74 个烷烃化合物临界温度 (CT_{EST}) 值与原实验观测值 (CT_{OBS}) 的相关图如图 4-3A 所示。

由 (4-9) 式对数模型估计 74 个烷烃化合物的临界压力 (CP_{EST}) 值与原实验观测值 CP_{OBS} 的相关情况如图 4-3B 所示。

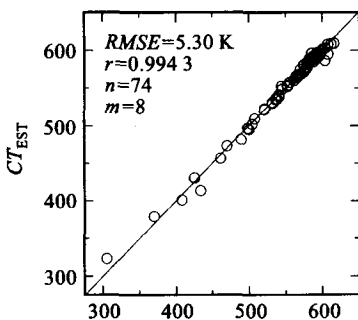


图 4-3A 74 个烷烃估计与实验
临界温度相关图

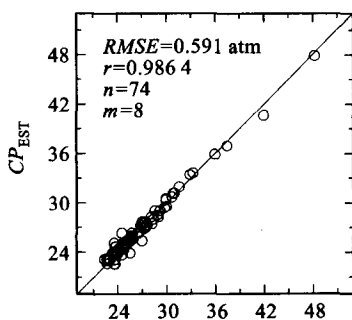


图 4-3B 74 个烷烃估计与实验
临界压力相关图

4.2.3 表面张力

68 个烷烃及其在 20°C 时的表面张力 (surface tensions 简称 ST) 选自 Needham 报告^[17]。实验表面张力参见表 4-6。由于第 10 个 MEDV-4 描述子 x_{10} 只有 2 个化合物不为零 即 68 号 2,2,3,3-四甲基戊烷 ($x_{10} = 1.0000$) 和 70 号 2,2,4,4-四甲基戊烷 ($x_{10} = 0.2500$) 第 9 个 MEDV-4 描述子 x_9 有 14 个非零样本。仿前述方法 分别建立 10 参数、去除 x_{10} 的 9 参数、去除 x_{10} 与 x_9 的 8 参数线性回归模型。结果表明 9 参数的对数模型 (4-10) 式具有最好的估计与预测表面张力的能力, 说明描述子 x_9 不能从模型中删去。

表 4-6 68 个烷烃的实验表面张力 ST (单位: dyn/cm, $1 \text{ dyn} = 10^{-5} \text{ N}$)

No	ST	No	ST	No	ST	No	ST	No	ST
6	16.00	21	20.44	35	18.77	50	22.01	64	22.80
7	15.00	22	18.76	36	21.56	51	22.80	65	21.77
9	18.42	23	21.76	37	21.14	52	21.77	66	23.22
10	17.38	24	20.60	38	21.52	53	22.01	67	23.27
11	18.12	25	21.17	39	21.99	54	22.81	68	23.38
12	16.30	26	21.00	41	22.92	55	22.81	69	21.98
13	17.37	27	19.60	42	21.88	56	21.86	70	20.37
14	20.26	28	20.99	43	22.34	57	20.51	71	23.31
15	19.29	29	20.05	44	22.34	58	20.04	72	22.38
16	19.79	30	19.73	45	20.80	59	22.41	73	23.87
17	18.02	31	20.63	46	22.34	60	22.80	74	22.80
18	19.96	32	21.64	47	21.30	61	21.27	75	23.75
19	18.15	33	21.51	48	21.30	62	21.17		
20	19.59	34	20.67	49	20.83	63	23.27		

$$\begin{aligned}\ln(36 - ST) = & (4.008\ 6 \pm 0.045\ 3) - (0.889\ 2 \pm 0.057\ 1)x_1 - \\ & (0.332\ 0 \pm 0.016\ 4)x_2 - (0.065\ 4 \pm 0.009\ 3)x_3 + \\ & (0.123\ 2 \pm 0.014\ 6)x_4 - (0.051\ 6 \pm 0.002\ 4)x_5 + \\ & (0.073\ 4 \pm 0.008\ 9)x_6 + (0.131\ 6 \pm 0.013\ 2)x_7 + \\ & (0.092\ 5 \pm 0.015\ 3)x_8 + (0.088\ 2 \pm 0.016\ 3)x_9, \\ & (4-10)\end{aligned}$$

$n = 68, m = 9, r = 0.991\ 4, RMSE = 0.260\ \text{dyn/cm}, F = 371.385$
(建模时)

$q = 0.986\ 4\ RMSV = 0.319\ \text{dyn/cm}$ (LOO 校验)

由(4-10)式获得的估计表面张力 (ST_{EST}) 及原实验值 (ST_{OBS}) 相关图见图 4-4。

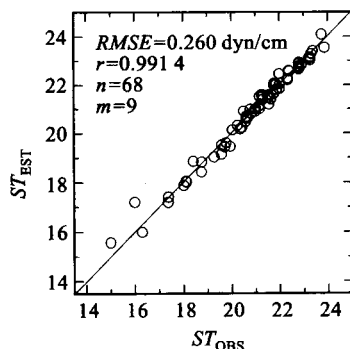


图 4-4 68 个烷烃估计与实验表面张力相关图

4.2.4 摩尔体积、摩尔折光率与蒸发热

选择 Needham 报告^[17]中 20 ℃时 69 个烷烃的摩尔体积 (molar volumes 简称 MV)、摩尔折光率 (molar refraction 简称 MR) 以及 25 ℃时的蒸发热 (heats of vaporization 简称 HV) 分别研究与 MEDV-4 分子描述子的相关性。这些性质的实验值参见表 4-7。

表 4-7 69 个烷烃的实验摩尔体积 MV (单位: cm^3/mol)、摩尔折光率 MR (单位: $1/\text{mol}$) 与蒸发热 HV (单位: kJ/mol)

No	MV	MR	HV	No	MV	MR	HV	No	MV	MR	HV
6	115.205	25.265 6	26.42	29	163.093	39.130 0	37.76	53	176.897	43.602 2	42.66
7	116.426	25.292 3	24.59	30	164.697	39.259 6	37.86	54	176.410	43.642 0	44.81
8	122.074	25.724 3	21.78	31	160.879	39.008 7	37.93	55	175.685	43.490 7	44.81
9	130.688	29.906 6	31.55	32	158.814	38.845 3	39.02	56	175.878	43.622 6	41.91
10	131.933	29.945 9	29.86	33	160.072	38.944 1	39.40	57	179.220	43.763 8	40.57
11	129.717	29.801 6	30.27	34	159.526	38.924 9	36.91	58	181.346	43.935 6	40.17
12	132.744	29.934 7	27.69	35	165.083	39.261 7	35.13	59	173.780	43.434 7	42.23
13	130.240	29.801 4	29.12	36	157.292	38.761 7	37.22	60	173.498	43.391 7	42.93
14	146.540	34.550 4	36.55	37	158.852	38.868 1	37.61	61	177.656	43.647 4	41.42
15	147.656	34.590 8	34.80	38	158.794	38.836 2	38.52	62	177.787	43.659 8	40.84
16	145.821	34.459 7	35.08	39	157.026	38.717 1	37.99	63	172.055	43.340 7	42.28
17	148.695	34.616 6	32.43	41	178.713	43.842 3	46.44	64	175.445	43.655 0	43.84
18	144.153	34.323 7	34.24	42	179.773	43.879 5	44.65	65	177.386	43.647 2	42.98
19	148.949	34.619 2	32.88	43	177.952	43.729 6	44.75	66	173.077	43.268 0	43.04
20	144.530	34.332 3	33.02	44	178.150	43.768 7	44.75	67	172.844	43.374 6	43.95
21	143.517	34.282 7	35.22	45	180.507	43.913 8	42.28	68	169.495	43.214 7	41.00
22	145.191	34.373 6	32.04	46	176.653	43.626 9	43.79	69	173.557	43.435 9	41.00
23	162.592	39.192 2	41.48	47	179.120	43.739 3	42.87	70	178.256	43.874 7	38.10
24	163.663	39.231 6	39.68	48	179.371	43.848 4	42.87	71	169.928	43.201 6	41.75
25	161.832	39.100 1	39.83	49	180.914	43.925 8	42.82	72	174.537	43.457 1	42.02
26	162.105	39.117 4	39.67	50	176.897	43.687 0	42.66	73	170.093	42.954 2	42.55
27	164.285	39.252 5	37.29	51	175.349	43.547 3	43.84	74	173.804	43.403 7	42.93
28	160.395	38.980 8	38.79	52	177.386	43.637 8	42.98	75	170.185	43.113 4	43.36

研究表明 除第 10 个 MEDV 描述子 x_{10} 外 9 个描述子的摩尔体积与 MEDV 模型具有最好的统计显著性及模型估计与预测能力。这是因为这 69 个烷烃中 MEDV 描述子 x_{10} 只有 2 个非零样本值。

$$MV = (89.182\ 9 \pm 0.929\ 5) - (21.715\ 0 \pm 1.283\ 4)x_1 + \\ (1.029\ 8 \pm 0.332\ 2)x_2 + (10.359\ 9 \pm 0.169\ 8)x_3 + \\ (16.246\ 2 \pm 0.318\ 8)x_4 + (10.837\ 0 \pm 0.052\ 1)x_5 + \\ (12.963\ 4 \pm 0.179\ 0)x_6 + (12.641\ 0 \pm 0.269\ 9)x_7 + \\ (10.511\ 0 \pm 0.318\ 2)x_8 + (7.153\ 6 \pm 0.348\ 0)x_9 \quad (4-11)$$

$n = 69, m = 9, r = 0.999\ 8, RMSE = 0.359\ 2\ \text{cm}^3/\text{mol}, F = 14\ 902.3$ (建模时)

$$q = 0.999\ 4, RMSV = 0.611\ 1\ \text{cm}^3/\text{mol} \quad (\text{LOO 校验})$$

由 (4-11) 式估计的摩尔体积 MV_{EST} 与实验值 MV_{OBS} 相关图见图 4-5A。

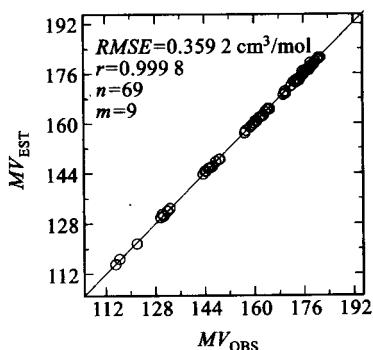


图 4-5A 69 个烷烃估计与实验摩尔体积相关图

对于摩尔折光率，也是 9 个 MEDV 描述子的 QSPR 模型 (4-12) 具有最好统计显著性及模型估计与预测能力。

$$\begin{aligned}
 MR = & (10.310\ 7 \pm 0.177\ 5) + (1.786\ 1 \pm 0.245\ 0)x_1 + \\
 & (3.000\ 1 \pm 0.063\ 4)x_2 + (3.199\ 1 \pm 0.032\ 4)x_3 + \\
 & (3.194\ 6 \pm 0.060\ 9)x_4 + (3.053\ 5 \pm 0.009\ 9)x_5 + \\
 & (2.623\ 2 \pm 0.032\ 4)x_6 + (2.070\ 1 \pm 0.051\ 5)x_7 + \\
 & (1.844\ 6 \pm 0.060\ 8)x_8 + (1.029\ 9 \pm 0.066\ 4)x_9 \quad (4-12)
 \end{aligned}$$

$$n = 69, m = 9, r = 0.999\ 9, RMSE = 0.068\ 59\ \text{mol}^{-1}, F = 37\ 851.9$$

(建模时)

$$q = 0.999\ 9, RMSV = 0.083\ 79\ \text{mol}^{-1} \quad (\text{LOO 校验})$$

对于蒸发热, 其结果表明删除 MEDV 描述子 x_{10} 和 x_9 后的 8 参数模型 (4-13) 具有最好的统计显著性以及模型估计与预测能力。在这组数据中 MEDV 描述子 x_{10} 只有 2 个非零值 描述子 x_9 有 14 个非零值。

$$\begin{aligned}
 HV = & (3.507\ 3 \pm 0.301\ 8) + (9.136\ 1 \pm 0.666\ 2)x_1 + \\
 & (5.902\ 0 \pm 0.097\ 0)x_2 + (3.213\ 6 \pm 0.095\ 6)x_3 + \\
 & (1.159\ 7 \pm 0.192\ 5)x_4 + (3.116\ 9 \pm 0.059\ 0)x_5 + \\
 & (1.122\ 6 \pm 0.059\ 0)x_6 - (0.225\ 2 \pm 0.090\ 6)x_7 + \\
 & (0.464\ 1 \pm 0.095\ 4)x_8 \quad (4-13)
 \end{aligned}$$

$$n = 69, m = 8, r = 0.999\ 1, RMSE = 0.231\ \text{kJ/mol}, F = 369\ 5.79$$

(建模时)

$$q = 0.998\ 7\ RMSV = 0.268\ \text{kJ/mol} \quad (\text{LOO 校验})$$

由 (4-12) 式估计的摩尔折光率 (MR_{EST}) 与实验值 (MR_{OBS}) 相关图见图 4-5B。由 (4-13) 式估计的蒸发热 (HV_{EST}) 与实验值 (HV_{OBS}) 相关图见图 4-5C。

4.2.5 热容与热焓

从 Gakh 报告^[16]中选择 134 个烷烃化合物在 300 K 时的热容 (heat capacity 简称 HC 和热焓 (enthalpy 简称 EN) 具体取值参见表 4-8。这 134 个化合物中 第 10 和第 9 两个 MEDV-4 描述

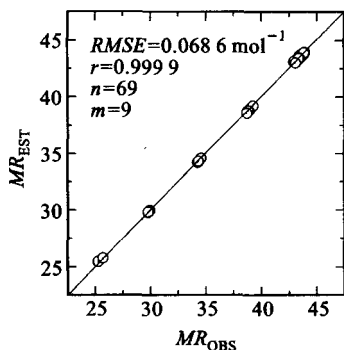
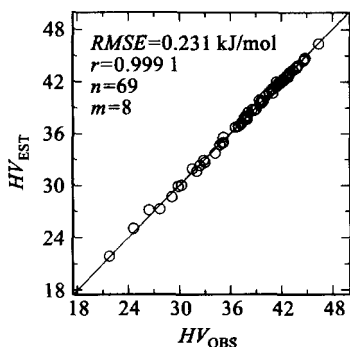


图 4-5B 69 个烷烃估计与实验 图 4-5C 69 个烷烃估计与实验
摩尔折光率相关图 蒸发热相关图



子的非零样本数较少, x_{10} 只有 4 个, 应当删去。 x_9 有 41 个, 大于 10% 能否删除 应考察 9 参数和 8 参数 QSPR 模型。

研究结果表明, 对于热容数据集, x_{10} 和 x_9 两个变量的引入虽然增加了相关系数 r 降低了均方根误差 $RMSE$ 但 F 统计量也同时减少了, 特别是各个回归系数的标准偏差都普遍偏高了。当进行 LOO 交互检验时, 以不含 x_{10} 和 x_9 的 8 参数模型为最优。这个模型如下:

$$\begin{aligned}
 HC = & (30.4631 \pm 3.3644) + (39.3769 \pm 6.7014)x_1 + \\
 & (20.2582 \pm 0.9897)x_2 + (14.3121 \pm 1.0470)x_3 + \\
 & (8.0909 \pm 1.9836)x_4 + (14.8974 \pm 0.3906)x_5 + \\
 & (11.0699 \pm 0.6175)x_6 + (8.1327 \pm 0.9504)x_7 + \\
 & (5.6535 \pm 0.8830)x_8
 \end{aligned} \quad (4-14)$$

$$n = 134, m = 8, r = 0.9882, RMSE = 3.871 \text{ J} \cdot \text{K}^{-1} \cdot \text{mol}^{-1}, F = 649.221 \quad (\text{建模时})$$

$$q = 0.9860, RMSV = 4.212 \text{ J} \cdot \text{K}^{-1} \cdot \text{mol}^{-1} \quad (\text{LOO 校验})$$

表 4-8 134 个烷烃的实验热容 HC (单位: $J \cdot K^{-1} \cdot mol^{-1}$) 与热焓 EN (单位: $J \cdot K^{-1}$)

No	HC	EN	No	HC	EN	No	HC	EN	No	HC	EN
10	143.01	26.61	28	185.18	33.05	47	217.10	38.16	65	219.70	39.25
11	140.88	26.32	29	193.35	33.76	48	208.20	37.53	66	214.10	37.36
12	142.26	25.40	30	186.52	33.39	49	210.40	37.99	67	215.20	38.07
13	142.21	24.77	31	191.96	33.43	50	214.00	38.20	68	213.34	35.86
14	166.00	33.56	32	182.72	32.47	51	206.80	37.02	69	208.50	35.06
15	165.40	31.21	33	190.58	36.07	52	214.60	38.07	70	215.77	36.44
16	164.50	30.71	34	186.77	32.13	53	217.20	37.53	71	219.50	36.23
17	167.70	29.50	35	189.45	32.55	54	213.00	40.71	72	205.00	36.15
18	161.80	28.62	36	188.20	32.17	55	214.30	40.50	73	213.40	36.61
19	171.70	29.58	37	192.72	32.55	56	209.90	36.61	74	209.00	36.07
20	166.70	29.33	38	193.05	34.31	57	210.70	36.61	75	217.86	38.37
21	166.80	31.84	39	189.07	33.26	58	209.10	36.86	81	235.10	43.43
22	164.20	28.28	40	188.28	31.84	59	213.30	36.28	82	230.50	42.43
23	188.70	38.12	42	210.90	40.42	60	214.00	36.86	83	239.40	42.76
24	188.20	35.82	43	209.70	39.92	61	212.50	36.02	84	231.80	41.92
25	186.82	35.31	44	210.40	39.71	62	213.50	36.44	85	231.90	42.09
26	188.03	35.06	45	212.40	38.83	63	210.50	35.98	86	233.20	42.58
27	189.33	34.23	46	207.70	37.82	64	216.10	38.70	87	237.10	42.80

续表

No	HC	EN	No	HC	EN	No	HC	EN	No	HC	EN
88	229.30	41.80	104	234.10	39.98	120	239.20	43.10	137	235.00	40.29
89	238.30	42.47	105	246.30	41.13	121	238.20	40.00	138	240.80	41.34
90	229.60	41.63	106	234.20	40.54	122	229.40	39.25	139	228.20	39.92
91	239.30	42.30	107	233.60	40.46	123	235.80	39.54	140	235.50	40.42
92	230.10	41.51	108	234.10	40.46	124	239.20	40.29	141	242.50	42.43
93	235.80	45.31	109	235.60	40.08	125	229.20	38.83	142	246.90	43.60
94	236.50	45.10	110	235.10	41.14	126	229.80	39.37	143	231.80	40.46
95	232.50	41.30	111	238.50	43.30	127	241.50	40.04	144	234.30	38.62
96	234.70	41.05	112	243.40	43.64	128	234.00	39.16	145	234.20	38.81
97	230.50	40.50	113	234.30	42.93	129	231.80	38.87	146	223.70	38.41
98	234.80	40.96	114	236.20	42.13	130	243.10	40.71	147	227.30	38.87
99	235.10	41.00	115	236.20	42.47	131	238.00	39.87	148	229.00	38.58
100	237.60	40.90	116	240.90	43.35	132	227.70	40.50	149	224.80	39.25
101	233.90	40.12	117	238.60	42.51	133	236.10	41.92	150	234.50	39.25
102	228.50	39.75	118	239.20	41.46	134	238.20	40.70			
103	238.90	40.54	119	237.70	44.85	135	243.00	42.43			

对热焓作类似研究的结果表明，虽然 10 参数方程的 $RMSE$ 无论是在建模阶段还是预测阶段都是三个模型中最小的，相关系数也是最大的，但是 F 也是最小的，这就相互矛盾。考察不含 x_{10} 的 9 参数模型 其 LOO 检验时的 $RMSV$ 与 10 参数模型差别不大 但建模时的 F 却比 10 参数模型有较大提高。考虑到 10 参数方程中各回归系数的相对标准偏差有四个大于 100%，同时考虑 x_{10} 的非零样本数少，因此，对于热焓数据集宜选 9 参数模型进行估计与预测。该模型如下：

$$\begin{aligned}
 EN = & (9.861\ 3 \pm 0.696\ 7) + (6.739\ 0 \pm 0.930\ 7)x_1 + \\
 & (2.600\ 7 \pm 0.242\ 4)x_2 + (1.495\ 1 \pm 0.155\ 0)x_3 + \\
 & (0.760\ 3 \pm 0.254\ 9)x_4 + (3.090\ 0 \pm 0.005\ 05)x_5 + \\
 & (2.343\ 5 \pm 0.133\ 3)x_6 + (1.567\ 5 \pm 0.199\ 9)x_7 + \\
 & (1.736\ 0 \pm 0.225\ 6)x_8 + (0.775\ 8 \pm 0.233\ 9)x_9 \quad (4-15)
 \end{aligned}$$

$n = 134, m = 9, r = 0.993\ 5, RMSE = 0.489\ J \cdot K^{-1}, F = 104\ 3.17$
 (建模时)

$$q = 0.992\ 0, RMSV = 0.541\ J \cdot K^{-1} \quad (\text{LOO 校验})$$

由式 (4-14) 估计的热容 HC_{EST} 与实验热容 HC_{OBS} 相关图见图 4-6A。由式 (4-15) 估计的热焓 (EN_{EST}) 与实验热焓 (EN_{OBS}) 相关图见图 4-6B。

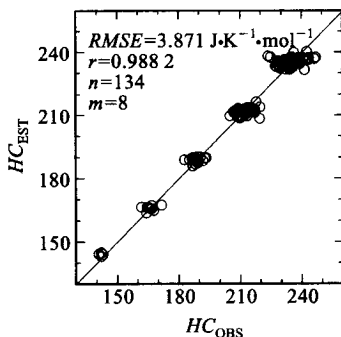


图 4-6A 134 个烷烃估计与实验热容相关图

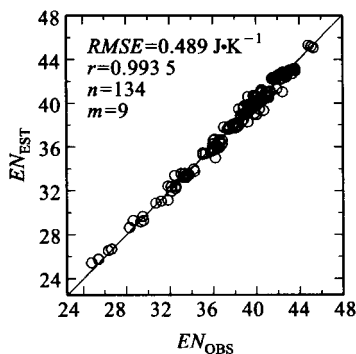


图 4-6B 134 个烷烃估计与实验热焓相关图

4.2.6 液态与气态生成焓

从 Somayajulu 报告^[18]中选择 62 个烷烃在 298.15 K 的液态生成焓 (heats of formation in liquid 简称 LF) 和气态生成焓 (heats of formation in gas 简称 GF) 具体取值参见表 4-9。在这 62 个烷烃的 MEDV 分子描述子中, x_{10}, x_9, x_8 分别只有 5, 14, 13 个非零样本。因而应分别考察 9, 8, 7 个参数的 QSPR 模型。

研究结果表明, 液态生成焓模型中以不含 x_{10} 的 9 参数方程为最优。结果如下:

$$\begin{aligned}
 LF = & (28.610\ 5 \pm 1.436\ 2) - (6.129\ 3 \pm 1.645\ 6)x_1 + \\
 & (1.590\ 1 \pm 0.511\ 7)x_2 + (4.663\ 5 \pm 0.201\ 6)x_3 + \\
 & (6.388\ 3 \pm 0.362\ 2)x_4 + (4.104\ 2 \pm 0.076\ 1)x_5 + \\
 & (4.455\ 1 \pm 0.290\ 8)x_6 + (4.428\ 2 \pm 0.393\ 1)x_7 + \\
 & (2.975\ 9 \pm 0.576\ 8)x_8 + (1.913\ 1 \pm 0.532\ 7)x_9, \quad (4-16)
 \end{aligned}$$

$n=62, m=9, r=0.997\ 7, RMSE=0.723\ \text{kJ/mol}, F=126\ 9.85$ (建模时)

$q=0.996\ 5, RMSV=0.893\ \text{kJ/mol}$ (LOO 校验)

气态生成焓模型也以 9 参数方程为最优。结果如下:

$$GF = (27.024\ 1 \pm 1.394\ 1) - (6.985\ 6 \pm 1.597\ 4)x_1 +$$

表 4-9 62 个烷烃的实验液态生成热 LF (单位: kJ/mol) 与气态生成热 GF (单位: kJ/mol)

No	LF	GF	No	LF	GF	No	LF	GF	No	LF	GF
2	22.50	20.24	18	54.83	46.65	34	61.44	52.61	63	66.33	56.42
3	28.79	24.82	19	56.17	48.30	35	61.97	53.57	68	66.54	57.07
4	35.34	30.15	20	56.07	48.17	36	60.63	51.73	69	66.40	56.81
5	36.95	32.15	21	53.77	45.34	37	60.98	51.97	70	66.95	58.16
6	41.40	35.00	22	56.63	48.96	38	59.69	50.48	71	66.46	56.67
7	42.95	36.92	23	59.74	49.82	39	60.46	51.38	72	65.18	55.37
8	45.61	40.27	24	60.98	51.50	40	62.40	53.99	74	64.47	54.48
9	47.52	39.96	25	60.34	50.82	41	65.64	54.54	75	65.85	55.81
10	48.82	41.66	26	60.17	50.69	44	66.82	56.19	76	71.95	59.67
11	48.28	41.02	27	62.63	53.71	45	68.88	58.83	77	74.04	62.17
12	51.00	44.35	28	60.40	51.16	56	67.57	57.70	80	73.58	61.80
13	49.48	42.49	29	61.47	52.44	57	67.60	58.12	108	72.84	62.22
14	53.63	44.89	30	62.26	53.21	58	70.11	60.53	121	72.53	62.08
15	54.93	46.60	31	61.58	52.61	59	67.18	57.31	126	77.32	67.29
16	54.35	45.96	32	60.23	50.91	61	67.88	59.97			
17	57.05	49.29	33	59.88	50.40	62	66.97	57.47			

$$\begin{aligned}
 & (0.334\ 7 \pm 0.496\ 7)x_2 + (3.826\ 5 \pm 0.195\ 7)x_3 + \\
 & (5.799\ 3 \pm 0.351\ 6)x_4 + (3.389\ 6 \pm 0.073\ 8)x_5 + \\
 & (4.207\ 8 \pm 0.282\ 2)x_6 + (4.520\ 5 \pm 0.381\ 5)x_7 + \\
 & (2.866\ 7 \pm 0.559\ 9)x_8 + (1.799\ 2 \pm 0.517\ 1)x_9 \quad (4-17)
 \end{aligned}$$

$n=62, m=9, r=0.997\ 0, RMSE=0.702\ \text{kJ/mol}, F=961.372$ (建模时)
 $q=0.995\ 4, RMSV=0.870\ \text{kJ/mol}$ (LOO 校验)

由式 4-16)估计的液态生成焓 (LF_{EST}) 与实验液态生成焓 (LF_{OBS}) 相关情况见图 4-7A。由式 (4-17)估计的气态生成焓 (GF_{EST}) 与实验气态生成焓 (GF_{OBS}) 相关情况见图 4-7B。

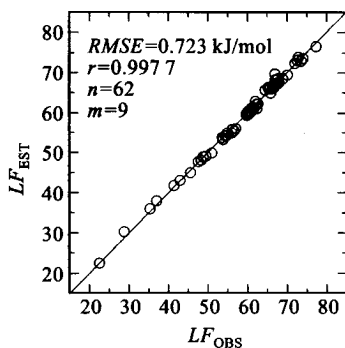


图 4-7A 62 个烷烃估计与实验
液态生成焓相关图

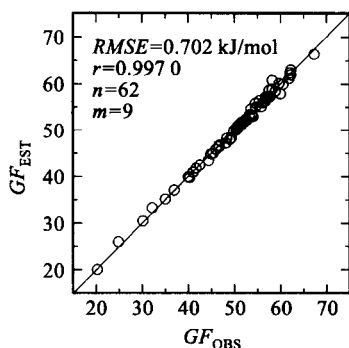


图 4-7B 62 个烷烃估计与实验
气态生成焓相关图

4.2.7 色谱保留指数

色谱保留指数是鉴定与识别化合物的最重要色谱参数之一。它与分子描述子之间的定量关系已有报道^[19~22]。选择武杰与陆婉珍等报告^[23]中的 64 个烷烃及其程序升温气相色谱保留指数为数据集, 实验值参见表 4-10。在所有 64 个烷烃中 MEDV 描述子 x_{10} 的值都为零, 因而实际上只有 9 个 MEDV 描述子 据此我们建立了质量非常高的 9 参数定量结构-保留指数相关 (quantitative structure-retention index relationship 简称 QSRR 模型^[24]。

表 4-10 64个烷烃的实验气相色谱保留指数 (GC-RI)

No GC-RI	No GC-RI	No GC-RI	No GC-RI	No GC-RI	No GC-RI
3 300.0	15 662.9	26 756.5	41 900.0	52 837.6	65 827.3
4 400.0	16 672.2	27 721.6	42 864.8	54 870.3	67 856.3
5 354.2	17 620.5	28 757.9	43 871.4	56 823.1	76 1 000.0
6 500.0	18 665.0	29 731.7	44 863.7	57 791.6	77 964.0
7 466.1	19 625.8	30 729.7	45 818.9	58 785.5	78 970.5
9 600.0	20 650.5	31 739.2	46 856.3	59 839.7	79 961.6
10 562.0	21 683.0	32 767.6	47 823.1	60 850.2	80 959.8
11 578.6	22 631.4	33 775.0	48 835.5	61 814.9	96 889.9
12 528.5	23 800.0	35 688.0	49 830.1	62 812.2	108 913.2
13 557.7	24 764.1	37 747.6	50 839.7	63 847.0	
14 700.0	25 772.1	38 759.7	51 858.8	64 847.0	

$$\begin{aligned}
 RI = & -(327.196\ 1 \pm 88.069\ 8) + (571.550\ 9 \pm 96.178\ 6)x_1 + \\
 & (244.245\ 8 \pm 30.985\ 6)x_2 + (85.202\ 6 \pm 6.192\ 0)x_3 - \\
 & (34.041\ 0 \pm 15.093\ 2)x_4 + (60.417\ 3 \pm 1.070\ 0)x_5 - \\
 & (35.828\ 5 \pm 14.883\ 3)x_6 - (92.060\ 8 \pm 24.024\ 0)x_7 - \\
 & (87.065\ 0 \pm 26.529\ 3)x_8 - (118.854\ 6 \pm 34.531\ 5)x_9 \quad (4-18)
 \end{aligned}$$

$n=64, m=9, r=0.999\ 2, RMSE=5.94, F=3\ 701.86$ (建模时)
 $q=0.998\ 8, RMSV=7.13$ (LOO校验)

由式 (4-18) 估计的气相色谱保留指数 RI_{EST} 与实验气相色谱保留指数 RI_{OBS} 总体相关情况如图 4-8。

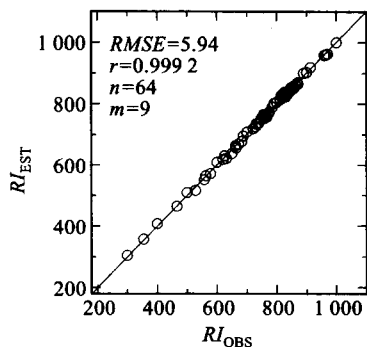


图 4-8 64 个烷烃估计与实验
保留指数相关图

4.3 链烃及醇结构表征与沸点相关

烷烃除氢原子外，非氢原子只有碳，而且所有化学键都是碳—碳单键，是最简单的有机化合物。为了说明 MEDV-4 可适用于更广泛的有机分子结构描述与表征，本节及下节将说明 MEDV-4 也能扩展应用到不饱和分子及含杂原子等比较复杂的分子体系，比如烯烃、炔烃、芳烃与多环芳烃等。在这些分子体系中，非氢原子除有碳原子外，还有氮、氧、硫等杂原子，原子属性也不像烷烃只有 C1 而有 C1, C2, CC, C3, N1, N2, O1, OO 等多种原子属性，因而构成多种不同性质的化学键，碳—碳单键、双键、三键、共轭碳—碳双键、碳—氧单键等，其相对键长随化学键的不同而变化。非氢原子相对电性也不再全等于 1。原子类型也不只限于有碳的原子类型。这里应当强调指出，我们定义的原子类型是以相连的非氢原子个数为依据，而不是以化学键的多重性来定义的，这样苯分子中的 6 个碳原子都属于第 2 原子类型。下面将分别讨论 MEDV-4 矢量在烯烃、炔烃、脂肪醇、取代芳烃、多环芳烃等不饱和烃及含杂原子体系的分子结构表征与相关性研究中的应用。

4.3.1 63个烯烃

从 Wessel 报告^[12]中选择 63 个链烯烃化合物及其沸点为数据集 (参见表 4-11)。在链烯烃中, 非氢原子仍只有碳原子, 与烷烃不同的是原子属性增加到 C1, C2, CC 等三种即化学键增加了碳-碳双键和共轭碳-碳双键。原子类型仍是 4 类。

63 个烯烃中碳原子数为 2~20 个, 其沸点从最小的乙烯 169.4 K 到 20 碳烯的 615.5 K。根据公式 2-1 计算各个 MEDV-4 矢量描述子, 有关结果见表 4-11。计算结果表明, 这些烯烃的 MEDV 分子描述子分布是不均匀的, 其中 x_{10} 没有非零样本; x_7, x_9 只有 3 个样本具有非零值 ($x_{7(17)} = 1.000\ 0, x_{7(18)} = 1.000\ 0, x_{7(29)} = 1.000\ 0; x_{9(17)} = 0.250\ 0, x_{9(18)} = 0.285\ 9, x_{9(55)} = 1.000\ 0$ 。下标括号内数值为样本序号); x_4 和 x_8 也分别只有 4 个和 6 个样本非零值 (均小于样本数的 10%); x_3, x_6 较多, 有 28 和 23 个非零样本; x_1, x_3, x_5 最多 分别有 63, 57 和 49 个非零样本。除顺反异构体外, MEDV 对烯烃有很好的分辨率 ($R_D = 100\%$)。

为了说明 MEDV 对烯烃分子结构的分辨率, 计算了 63 个烯烃两两之间的 1 953 个模式距离, 结果所有模式距离都不为零, 这说明 MEDV 可以区分所有的 63 个样本。模式距离的最小值为 0.011 9 最大值为 26.003 3。距离最小的两个样本是 45 号(辛烯-3)与 46 号样本(辛烯-4), 这与它们结构之间的相似性是很好的。如果要表征顺反异构体, 可在 MEDV 方法中引入几何距离。本文之所以未引入几何距离, 是考虑到这些顺反异构体之间沸点的差别很小, 一般只有 1~2 K 有的小于 1 K 如顺式与反式 3-甲基庚烷只相差 0.1 K 前者为 368.9 K 后者为 368.8 K。全部在测定沸点的误差范围与估计范围之内。因此, 作者选择了全部顺式结构及其沸点, 当然也可选择全部的反式结构与沸点, 或者顺反异构体的平均值。

表 4-11 63 个烯烃的沸点 $b.p.$ (单位:K) 和 7 个 MEDV-4 向量分子描述子值

No	化合物	x_1	x_2	x_3	x_4	x_5	x_6	x_8	$b.p.$
1	propylene	0.285 9	2.320 8	0	0	0	0	0	225.4
2	1,3-butadiene	0.127 1	2.859 3	0	0	1.143 7	0	0	268.7
3	1-butene	0.121 4	2.856 7	0	0	1.000 0	0	0	266.9
4	cis-2-butene	0.121 4	2.571 9	0	0	1.320 8	0	0	276.9
5	isobutene	0.821 9	0	3.320 8	0	0	0	0	266.2
6	isoprene	0.515 5	1.696 7	2.429 6	0	0	1.143 7	0	307.2
7	1-pentene	0.066 8	3.089 2	0	0	2.250 0	0	0	301.1
8	2,3-dimethyl-1-butene	1.000 9	0	5.356 7	0	0	0	1.000 0	328.8
9	2,3-dimethyl-2-butene	0.985 6	0	5.143 7	0	0	0	1.320 8	346.3
10	2-ethyl-1-butene	0.305 3	2.794 1	1.820 8	0	0.250 0	2.000 0	0	337.8
11	1-hexene	0.042 2	3.218 5	0	0	3.611 1	0	0	336.6
12	cis-2-hexene	0.042 2	2.902 0	0	0	3.978 1	0	0	342.0
13	2-methyl-1-pentene	0.415 2	2.018 4	2.431 9	0	1.000 0	1.250 0	0	335.2
14	4-methyl-1-pentene	0.383 5	2.328 9	2.121 4	0	1.000 0	1.250 0	0	327.0
15	1-heptene	0.029 0	3.300 6	0	0	5.034 7	0	0	366.8
16	1-octene	0.021 2	3.357 4	0	0	6.498 3	0	0	394.4

续表

No	化合物	x ₁	x ₂	x ₃	x ₄	x ₅	x ₆	x ₈	b, p.
17	2,4,4-trimethyl-1-pentene	1.423 7	1.285 9	2.654 1	3.232 5	0	1.000 0	0	374.6
18	2,4,4-trimethyl-2-pentene	1.400 6	1.321 9	2.364 2	3.242 8	0	1.320 8	0	378.1
19	1-decene	0.012 7	3.430 8	0	0	9.501 5	0	0	443.8
20	1-dodecene	0.008 5	3.476 1	0	0	12.568 7	0	0	486.5
21	1-tetradecene	0.006 0	3.506 9	0	0	15.676 5	0	0	524.2
22	1-hexadecene	0.004 5	3.529 2	0	0	18.812 4	0	0	558.0
23	1-octadecene	0.003 5	3.546 0	0	0	21.968 8	0	0	588.0
24	2-methyl-1-butene	0.518 4	1.535 9	2.570 8	0	0	1.000 0	0	304.3
25	2-methyl-2-butene	0.492 8	1.571 9	2.285 9	0	0	1.320 8	0	311.7
26	3-methyl-1-butene	0.492 8	1.820 8	2.285 9	0	0	1.000 0	0	293.2
27	2,3-dimethyl-1,3-butadiene	1.020 1	0	5.393 4	0	0	0	1.143 7	341.9
28	1,5-hexadiene	0.044 5	3.589 7	0	0	3.611 1	0	0	332.6
29	3,3-dimethyl-1-butene	1.114 2	2.070 8	0	3.285 9	0	0	0	314.4
30	2-methyl-2-pentene	0.383 5	2.064 6	2.121 4	0	1.000 0	1.606 7	0	340.4
31	3-methyl-1-pentene	0.299 3	3.053 3	1.535 9	0	0.250 0	2.000 0	0	327.3
32	ethylene	1.320 8	0	0	0	0	0	0	169.4

续表

No	化合物	x_1	x_2	x_3	x_4	x_5	x_6	x_8	$b.p.$
33	1 - nonene	0.016 1	3.399 0	0	0	7.989 7	0	0	420.0
34	1 - undecene	0.010 3	3.455 9	0	0	11.028 9	0	0	465.8
35	1 - tridecene	0.007 1	3.492 9	0	0	14.118 5	0	0	505.9
36	1 - pendeene	0.005 2	3.518 9	0	0	17.241 5	0	0	541.6
37	1 - heptadecene	0.004 0	3.538 1	0	0	20.388 4	0	0	573.5
38	1 - nonadecene	0.003 1	3.553 0	0	0	23.553 2	0	0	602.2
39	1 - eicosene	0.002 8	3.559 2	0	0	25.141 0	0	0	615.5
40	<i>cis</i> - 2 - pentene	0.066 8	2.778 7	0	0	2.606 7	0	0	310.1
41	<i>cis</i> - 3 - hexene	0.042 2	2.876 3	0	0	4.014 0	0	0	339.6
42	<i>cis</i> - 2 - heptene	0.029 0	2.982 0	0	0	5.406 0	0	0	371.6
43	<i>cis</i> - 3 - heptene	0.029 0	2.950 4	0	0	5.452 2	0	0	368.9
44	<i>cis</i> - 2 - octene	0.021 2	3.037 9	0	0	6.871 8	0	0	398.8
45	<i>cis</i> - 3 - octene	0.021 2	3.004 1	0	0	6.922 2	0	0	396.0
46	<i>cis</i> - 4 - octene	0.021 2	2.998 1	0	0	6.932 5	0	0	395.7
47	3 - methyl - <i>cis</i> - 2 - pentene	0.299 3	2.778 7	1.535 9	0	0.285 9	2.320 8	0	340.8
48	4 - methyl - <i>cis</i> - 2 - pentene	0.383 5	2.028 7	2.121 4	0	1.320 8	1.285 9	0	329.5

续表

No	化合物	x_1	x_2	x_3	x_4	x_5	x_6	x_8	$b, p.$
49	2-ethyl-1-hexene	0.368 1	2.258 8	2.383 3	0	2.250 0	1.361 1	0	365.0
50	3-methyl-1-hexene	0.226 1	3.432 6	1.397 0	0	1.361 1	2.250 0	0	357.0
51	4-methyl-1-hexene	0.220 0	3.458 2	1.371 4	0	1.361 1	2.250 0	0	359.9
52	2-methyl-1-heptene	0.342 7	2.403 5	2.360 8	0	3.611 1	1.423 6	0	392.4
53	2-ethyl-1-pentene	0.228 2	3.179 4	1.681 9	0	1.361 1	2.250 0	0	367.1
54	3-ethyl-1-pentene	0.196 0	4.008 0	0.785 9	0	0.750 0	3.000 0	0	357.3
55	2,3,3-trimethyl-1-butene	1.733 4	0	3.070 8	3.535 9	0	0	0	351.0
56	2,3-dimethyl-1-hexene	0.663 1	1.972 9	4.280 3	0	1.000 0	1.611 1	1.000 0	383.6
57	2-ethyl-1-hexene	0.191 3	3.374 7	1.633 3	0	2.673 6	2.361 1	0	393.1
58	cis-1,3-pentadiene	0.069 1	2.945 2	0	0	2.573 4	0	0	317.2
59	1,4-pentadiene	0.071 5	3.456 2	0	0	2.250 0	0	0	299.1
60	trans-1,4-hexadiene	0.044 5	3.273 7	0	0	3.9781	0	0	338.1
61	cis-2,4-hexadiene	0.043 3	2.915 0	0	0	4.130 1	0	0	356.6
62	2,5-dimethyl-1,5-hexadiene	0.740 7	1.536 9	4.900 1	0	1.000 0	2.500 0	0.111 1	387.4
63	2,5-dimethyl-2,4-hexadiene	0.673 2	1.553 8	4.276 3	0	1.143 7	2.859 3	0.127 1	407.6

应用 MLR 方法建立烯烃沸点与相应 MEDV 矢量描述子间的相关模型^[25]。研究表明, 直接以沸点为因变量, 建立 5 参数(不含 MEDV 描述子 x_4, x_7, x_8, x_9)模型、6 参数(不含描述子 x_4, x_7, x_8)模型、7 参数(不含 x_4, x_8)模型, 其估计均方根误差都在 12 K 以上, 不符合沸点测定的要求, 这样的估计没有应用价值。前面已经说明, 烷烃多种物理性质如沸点、密度、折光率等的对数模型更优。实验证明, 烯烃沸点也有同样的规律。随着碳原子数的增加或碳链的增长, 其沸点变化将趋于平缓, 故考察对数模型。研究表明, 7 参数与 9 参数对数模型具有良好统计性和良好估计与预测能力, 而 5 参数及 6 参数模型其估计 RMSE 值均大于 11 K 不能作为估计与预测模型, 说明 x_4, x_8 不能从模型中删除, 否则会影响 MEDV 对有关样本的分辨信息。进一步考察 7 参数与 9 参数模型结果可知, 在 9 参数模型中由于 32 号(乙烷)与 55 号(2,3,3-三甲基-1-丁烯)样本的预测误差特别巨大(分别为 -72.6 K 和 -1849.1 K) 因此导致 RMSE 达到 233.2 K, 如果不考虑这两个样本, 则 LOO 检验预测结果相当好, RMSE 可达 6 K 以下。而 7 参数模型仅 32 号比较大(-73.3 K), 所以误差稍小。从上述结果看, 7 参数模型更优一些, 其回归方程为:

$$\ln(801 - b.p.) = (6.2798 \pm 0.0237) + (0.1345 \pm 0.0202)x_1 + \\ (0.0058 \pm 0.0076)x_2 - (0.0356 \pm 0.0031)x_3 - \\ (0.0817 \pm 0.0051)x_4 - (0.0432 \pm 0.0003)x_5 - \\ (0.0587 \pm 0.0044)x_6 - (0.0751 \pm 0.0102)x_7 \quad (4-19) \\ n=63, m=7, r=0.9992, RMSE=4.43 \text{ K}, F=5206.72 \quad (\text{建模时}) \\ q=0.9976, RMSV=10.47 \text{ K} \quad (\text{LOO 校验})$$

由(4-19)估计的烯烃沸点(BP_{EST})及原实验观测沸点(BP_{OBS})相关情况见图 4-9。

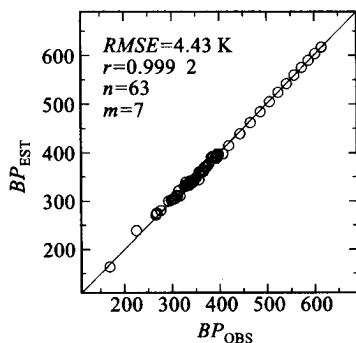


图 4-9 63 个烯烃估计与实验
烯烃沸点相关图

4.3.2 147 个链烃

为了说明 MEDV 矢量描述子也适用于表征烷烃、烯烃及炔烃化合物构成的混合体系，分别从 Wessel 等^[11, 12, 26]报告中选取非环链烃 147 个，即 69 个烷烃、61 个烯烃和 17 个炔烃等链烃化合物作为研究对象。这 147 个链烃化合物的实验沸点见表 4-12A，计算的 9 个 MEDV 分子描述子结果见表 4-12B。这 9 个 MEDV 描述子对这 147 个链烃分子结构的分辨率达 100%。应用 MLR 方法建立 147 个链烃分子的沸点与 MEDV 描述子的相关关系，研究结果表明，9 参数的对数模型质量最优：

$$\begin{aligned} \ln(801 - b.p.) = & (6.280\ 40 \pm 0.012\ 80) + (0.122\ 40 \pm 0.018\ 50)x_1 + \\ & (0.000\ 99 \pm 0.004\ 00)x_2 - (0.033\ 27 \pm 0.003\ 07)x_3 - \\ & (0.051\ 89 \pm 0.004\ 97)x_4 - (0.042\ 35 \pm 0.000\ 25)x_5 - \\ & (0.053\ 76 \pm 0.001\ 96)x_6 - (0.066\ 33 \pm 0.003\ 31)x_7 - \\ & (0.073\ 45 \pm 0.004\ 21)x_8 - (0.098\ 30 \pm 0.005\ 99)x_9 \end{aligned} \quad (4-20)$$

$n = 147, m = 9, r = 0.9985, RMSE = 4.26 \text{ K}, F = 5083.78$ (建模时)

$q = 0.9983, RMSV = 4.51 \text{ K}$ (LOO 预测)

以上结果充分说明, 模型 (4-20) 式具有良好的估计与预测能力。由 (4-20) 式估计的链烃沸点 (BP_{EST}) 及原实验沸点 (BP_{OBS}) 相关情况如图 4-10。

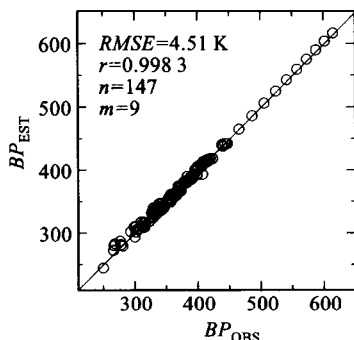


图 4-10 147 个链烃估计与实验
链烃沸点相关图

4.3.3 106 个醇

从李志良等的报告^[27]中, 选择 106 个脂肪醇及其实验沸点 (表 4-13A) 其沸点范围从 64.60°C 到 245.00°C 碳原子数从 1 到 11 个。这些脂肪醇的分子结构由 MEDV-4 矢量描述子表征 (MEDV 描述子计算结果见表 4-13B)^[28], 其中甲醇分子 (No.1) 的 MEDV-4 矢量中只有描述子 x_1 具有非零样本值, 所有 106 个醇中只有两个化合物即 2,3,3-三甲基-2-丁醇 (No.44) 和 2,2,3-三甲基-3-己醇 (No.90) 具有非零的描述子 x_{10} 。因此, 应考虑建立 10 参数与 9 参数模型。

表 4-12A 147 个链烃分子与实验沸点 $b.p.$ (单位: K)

No	化合物	$b.p.$	No	化合物	$b.p.$
1	1,3 - butadiene	268.7	20	1 - tetradecene	524.2
2	1 - butene	266.9	21	1 - hexadecene	558.0
3	<i>cis</i> - 2 - butene	276.9	22	1 - octadecene	588.0
4	isobutene	266.2	23	2 - methyl - 1 - butene	304.3
5	isoprene	307.2	24	2 - methyl - 2 - butene	311.7
6	1 - pentene	301.1	25	3 - methyl - 1 - butene	293.2
7	2,3 - dimethyl - 1 - butene	328.8	26	2,3 - dimethyl - 1,3 - butadiene	341.9
8	2,3 - dimethyl - 2 - butene	346.3	27	1,5 - hexadiene	332.6
9	2 - ethyl - 1 - butene	337.8	28	3,3 - dimethyl - 1 - butene	314.4
10	1 - hexene	336.6	29	2 - methyl - 2 - pentene	340.4
11	<i>cis</i> - 2 - hexene	342.0	30	3 - methyl - 1 - pentene	327.3
12	2 - methyl - 1 - pentene	335.2	31	1 - nonene	420.0
13	4 - methyl - 1 - pentene	327.0	32	1 - undecene	465.8
14	1 - heptene	366.8	33	1 - tridecene	505.9
15	1 - octene	394.4	34	1 - pentdecene	541.6
16	2,4,4 - trimethyl - 1 - pentene	374.6	35	1 - heptadecene	573.5
17	2,4,4 - trimethyl - 2 - pentene	378.1	36	1 - nonadecene	602.2
18	1 - decene	443.8	37	1 - eicosene	615.5
19	1 - dodecene	486.5	38	<i>cis</i> - 2 - pentene	310.1

续表

No	化合物	b. p.	No	化合物	b. p.
39	cis - 3 - hexene	339.6	58	trans - 1,4 - hexadiene	338.1
40	cis - 2 - heptene	371.6	59	cis - 2,4 - hexadiene	356.6
41	cis - 3 - heptene	368.9	60	2,5 - dimethyl - 1,5 - hexadiene	387.4
42	cis - 2 - octene	398.8	61	2,5 - dimethyl - 2,4 - hexadiene	407.6
43	cis - 3 - octene	396.0	62	hexane	342.1
44	cis - 4 - octene	395.7	63	2 - methylpentane	333.4
45	3 - methyl - cis - 2 - pentene	340.8	64	3 - methylpentane	336.4
46	4 - methyl - cis - 2 - pentene	329.5	65	2,2 - dimethylbutane	322.8
47	2 - methyl - 1 - hexene	365.0	66	2,3 - dimethylbutane	331.1
48	3 - methyl - 1 - hexene	357.0	67	heptane	371.5
49	4 - methyl - 1 - hexene	359.9	68	2 - methylhexane	363.1
50	2 - methyl - 1 - heptene	392.4	69	3 - methylhexane	365.1
51	2 - ethyl - 1 - pentene	367.1	70	2,2 - dimethylpentane	352.3
52	3 - ethyl - 1 - pentene	357.3	71	2,3 - dimethylpentane	362.9
53	2,3,3 - trimethyl - 1 - butene	351.0	72	2,4 - dimethylpentane	353.6
54	2,3 - dimethyl - 1 - hexene	383.6	73	3,3 - dimethylpentane	359.2
55	2 - ethyl - 1 - hexene	393.1	74	3 - ethylpentane	366.6
56	cis - 1,3 - pentadiene	317.2	75	2,2,3 - trimethylbutane	354.0
57	1,4 - pentadiene	299.1	76	octane	398.8

续表

No	化合物	<i>b. p.</i>	No	化合物	<i>b. p.</i>
77	2 - methylheptane	390.7	96	4 - methyloctane	415.5
78	3 - methylheptane	391.1	97	2,2 - dimethylheptane	405.8
79	4 - methylheptane	390.8	98	2,3 - dimethylheptane	413.6
80	2,2 - dimethylhexane	379.9	99	2,4 - dimethylheptane	406.6
81	2,3 - dimethylhexane	388.7	100	2,5 - dimethylheptane	409.1
82	2,4 - dimethylhexane	382.5	101	2,6 - dimethylheptane	408.3
83	2,5 - dimethylhexane	382.1	102	3,3 - dimethylheptane	410.4
84	3,3 - dimethylhexane	385.1	103	3,4 - dimethylheptane	413.2
85	3,4 - dimethylhexane	390.8	104	3,5 - dimethylheptane	409.1
86	3 - ethylhexane	391.6	105	4,4 - dimethylheptane	408.3
87	2,2,3 - trimethylpentane	383.1	106	3 - ethylheptane	416.1
88	2,2,4 - trimethylpentane	372.3	107	4 - ethylheptane	414.3
89	2,3,3 - trimethylpentane	387.8	108	2,2,3 - trimethylhexane	404.8
90	2,3,4 - trimethylpentane	386.5	109	2,2,4 - trimethylhexane	399.6
91	2 - methyl - 3 - ethylpentane	388.7	110	2,2,5 - trimethylhexane	397.1
92	3 - methyl - 3 - ethylpentane	391.3	111	2,3,3 - trimethylhexane	410.8
93	nonane	423.8	112	2,3,4 - trimethylhexane	412.1
94	2 - methyloctane	415.9	113	2,3,5 - trimethylhexane	404.4
95	3 - methyloctane	416.9	114	2,4,4 - trimethylhexane	399.6

续表

No	化合物	<i>b. p.</i>	No	化合物	<i>b. p.</i>
115	3,3,4 - trimethylhexane	413.6	132	vinylacetylene	278.2
116	2 - methyl - 3 - ethylhexane	411.1	133	dimethylacetylene	300.1
117	2 - methyl - 4 - ethylhexane	406.9	134	ethylacetylene	281.2
118	3 - methyl - 3 - ethylhexane	413.7	135	2 - methyl - 1 - butene - 3 - yne	305.4
119	3 - methyl - 4 - ethylhexane	413.5	136	1 - pentene - 3 - yne	332.4
120	2,2,3,4 - tetramethylpentane	406.1	137	1 - pentene - 4 - yne	315.6
121	2,3,3,4 - tetramethylpentane	414.6	138	3 - methyl - 1 - butyne	302.1
122	2,2 - dimethyl - 3 - ethylpentane	406.9	139	1 - pentyne	313.3
123	2,3 - dimethyl - 3 - ethylpentane	414.7	140	2 - pentyne	329.3
124	2,4 - dimethyl - 3 - ethylpentane	409.8	141	1 - hexyne	344.5
125	3,3 - diethylpentane	419.3	142	2 - hexyne	357.7
126	decane	447.2	143	3 - hexyne	354.3
127	2 - methylnonane	440.1	144	1 - heptyne	372.9
128	3 - methylnonane	440.9	145	1 - octyne	399.3
129	4 - methylnonane	438.8	146	1 - nonyne	424.0
130	5 - methylnonane	438.2	147	1 - decyne	447.2
131	methylacetylene	249.9			

表 4-12B 147 个链烃的 MEDV-4 向量描述子

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1	0.127 1	2.859 3	0.000 0	0.000 0	1.143 7	0.000 0	0.000 0	0.000 0	0.000 0
2	0.121 4	2.856 7	0.000 0	0.000 0	1.000 0	0.000 0	0.000 0	0.000 0	0.000 0
3	0.121 4	2.571 9	0.000 0	0.000 0	1.320 8	0.000 0	0.000 0	0.000 0	0.000 0
4	0.821 9	0.000 0	3.320 8	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0
5	0.515 5	1.696 7	2.429 6	0.000 0	0.000 0	1.143 7	0.000 0	0.000 0	0.000 0
6	0.066 8	3.089 2	0.000 0	0.000 0	2.250 0	0.000 0	0.000 0	0.000 0	0.000 0
7	1.000 9	0.000 0	5.356 7	0.000 0	0.000 0	0.000 0	0.000 0	1.000 0	0.000 0
8	0.985 6	0.000 0	5.143 7	0.000 0	0.000 0	0.000 0	0.000 0	1.320 8	0.000 0
9	0.305 3	2.794 1	1.820 8	0.000 0	0.250 0	2.000 0	0.000 0	0.000 0	0.000 0
10	0.042 2	3.218 5	0.000 0	0.000 0	3.611 1	0.000 0	0.000 0	0.000 0	0.000 0
11	0.042 2	2.902 0	0.000 0	0.000 0	3.978 1	0.000 0	0.000 0	0.000 0	0.000 0
12	0.415 2	2.018 4	2.431 9	0.000 0	1.000 0	1.250 0	0.000 0	0.000 0	0.000 0
13	0.383 5	2.328 9	2.121 4	0.000 0	1.000 0	1.250 0	0.000 0	0.000 0	0.000 0
14	0.029 0	3.300 6	0.000 0	0.000 0	5.034 7	0.000 0	0.000 0	0.000 0	0.000 0
15	0.021 2	3.357 4	0.000 0	0.000 0	6.498 3	0.000 0	0.000 0	0.000 0	0.000 0
16	1.423 7	1.285 9	2.654 1	3.232 5	0.000 0	1.000 0	1.000 0	0.000 0	0.250 0

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
17	1.400 6	1.321 9	2.364 2	3.242 8	0.000 0	1.320 8	1.000 0	0.000 0	0.285 9
18	0.012 7	3.430 8	0.000 0	0.000 0	9.501 5	0.000 0	0.000 0	0.000 0	0.000 0
19	0.008 5	3.476 1	0.000 0	0.000 0	12.568 7	0.000 0	0.000 0	0.000 0	0.000 0
20	0.006 0	3.506 9	0.000 0	0.000 0	15.676 5	0.000 0	0.000 0	0.000 0	0.000 0
21	0.004 5	3.529 2	0.000 0	0.000 0	18.812 4	0.000 0	0.000 0	0.000 0	0.000 0
22	0.003 5	3.546 0	0.000 0	0.000 0	21.968 8	0.000 0	0.000 0	0.000 0	0.000 0
23	0.518 4	1.535 9	2.570 8	0.000 0	0.000 0	1.000 0	0.000 0	0.000 0	0.000 0
24	0.492 8	1.571 9	2.285 9	0.000 0	0.000 0	1.320 8	0.000 0	0.000 0	0.000 0
25	0.492 8	1.820 8	2.285 9	0.000 0	0.000 0	1.000 0	0.000 0	0.000 0	0.000 0
26	1.020 1	0.000 0	5.393 4	0.000 0	0.000 0	0.000 0	0.000 0	1.143 7	0.000 0
27	0.044 5	3.589 7	0.000 0	0.000 0	3.611 1	0.000 0	0.000 0	0.000 0	0.000 0
28	1.114 2	2.070 8	0.000 0	3.285 9	0.000 0	0.000 0	1.000 0	0.000 0	0.000 0
29	0.383 5	2.064 6	2.121 4	0.000 0	1.000 0	1.606 7	0.000 0	0.000 0	0.000 0
30	0.299 3	3.053 3	1.535 9	0.000 0	0.250 0	2.000 0	0.000 0	0.000 0	0.000 0
31	0.016 1	3.399 0	0.000 0	0.000 0	7.989 7	0.000 0	0.000 0	0.000 0	0.000 0
32	0.010 3	3.455 9	0.000 0	0.000 0	11.028 9	0.000 0	0.000 0	0.000 0	0.000 0

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
33	0.007 1	3.492 9	0.000 0	0.000 0	14.118 5	0.000 0	0.000 0	0.000 0	0.000 0
34	0.005 2	3.518 9	0.000 0	0.000 0	17.241 5	0.000 0	0.000 0	0.000 0	0.000 0
35	0.004 0	3.538 1	0.000 0	0.000 0	20.388 4	0.000 0	0.000 0	0.000 0	0.000 0
36	0.003 1	3.553 0	0.000 0	0.000 0	23.553 2	0.000 0	0.000 0	0.000 0	0.000 0
37	0.002 8	3.559 2	0.000 0	0.000 0	25.141 0	0.000 0	0.000 0	0.000 0	0.000 0
38	0.066 8	2.778 7	0.000 0	0.000 0	2.606 7	0.000 0	0.000 0	0.000 0	0.000 0
39	0.042 2	2.876 3	0.000 0	0.000 0	4.014 0	0.000 0	0.000 0	0.000 0	0.000 0
40	0.029 0	2.982 0	0.000 0	0.000 0	5.406 0	0.000 0	0.000 0	0.000 0	0.000 0
41	0.029 0	2.950 4	0.000 0	0.000 0	5.452 2	0.000 0	0.000 0	0.000 0	0.000 0
42	0.021 2	3.037 9	0.000 0	0.000 0	6.871 8	0.000 0	0.000 0	0.000 0	0.000 0
43	0.021 2	3.004 1	0.000 0	0.000 0	6.922 2	0.000 0	0.000 0	0.000 0	0.000 0
44	0.021 2	2.998 1	0.000 0	0.000 0	6.932 5	0.000 0	0.000 0	0.000 0	0.000 0
45	0.299 3	2.778 7	1.535 9	0.000 0	0.285 9	2.320 8	0.000 0	0.000 0	0.000 0
46	0.383 5	2.028 7	2.121 4	0.000 0	1.320 8	1.285 9	0.000 0	0.000 0	0.000 0
47	0.368 1	2.258 8	2.383 3	0.000 0	2.250 0	1.361 1	0.000 0	0.000 0	0.000 0
48	0.226 1	3.432 6	1.397 0	0.000 0	1.361 1	2.250 0	0.000 0	0.000 0	0.000 0

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
49	0.220 0	3.458 2	1.371 4	0.000 0	1.361 1	2.250 0	0.000 0	0.000 0	0.000 0
50	0.342 7	2.403 5	2.360 8	0.000 0	3.611 1	1.423 6	0.000 0	0.000 0	0.000 0
51	0.228 2	3.179 4	1.681 9	0.000 0	1.361 1	2.250 0	0.000 0	0.000 0	0.000 0
52	0.196 0	4.008 0	0.785 9	0.000 0	0.750 0	3.000 0	0.000 0	0.000 0	0.000 0
53	1.733 4	0.000 0	3.070 8	3.535 9	0.000 0	0.000 0	0.000 0	0.000 0	1.000 0
54	0.663 1	1.972 9	4.280 3	0.000 0	1.000 0	1.611 1	0.000 0	1.000 0	0.000 0
55	0.191 3	3.374 7	1.633 3	0.000 0	2.673 6	2.361 1	0.000 0	0.000 0	0.000 0
56	0.069 1	2.945 2	0.000 0	0.000 0	2.573 4	0.000 0	0.000 0	0.000 0	0.000 0
57	0.071 5	3.456 2	0.000 0	0.000 0	2.250 0	0.000 0	0.000 0	0.000 0	0.000 0
58	0.044 5	3.273 7	0.000 0	0.000 0	3.978 1	0.000 0	0.000 0	0.000 0	0.000 0
59	0.043 3	2.915 0	0.000 0	0.000 0	4.130 1	0.000 0	0.000 0	0.000 0	0.000 0
60	0.740 7	1.536 9	4.900 1	0.000 0	1.000 0	2.500 0	0.000 0	0.111 1	0.000 0
61	0.673 2	1.553 8	4.276 3	0.000 0	1.143 7	2.859 3	0.000 0	0.127 1	0.000 0
62	0.040 0	2.847 2	0.000 0	0.000 0	3.611 1	0.000 0	0.000 0	0.000 0	0.000 0
63	0.375 0	1.972 2	2.111 1	0.000 0	1.000 0	1.250 0	0.000 0	0.000 0	0.000 0
64	0.284 7	2.722 2	1.500 0	0.000 0	0.250 0	2.000 0	0.000 0	0.000 0	0.000 0

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
65	1.083 3	1.750 0	0.000 0	3.250 0	0.000 0	0.000 0	1.000 0	0.000 0	0.000 0
66	0.944 4	0.000 0	5.000 0	0.000 0	0.000 0	0.000 0	0.000 0	1.000 0	0.000 0
67	0.027 8	2.927 2	0.000 0	0.000 0	5.034 7	0.000 0	0.000 0	0.000 0	0.000 0
68	0.330 0	2.208 3	2.062 5	0.000 0	2.250 0	1.361 1	0.000 0	0.000 0	0.000 0
69	0.213 6	3.097 2	1.361 1	0.000 0	1.361 1	2.250 0	0.000 0	0.000 0	0.000 0
70	0.937 5	2.333 3	0.000 0	3.111 1	1.000 0	0.000 0	1.250 0	0.000 0	0.000 0
71	0.708 3	1.472 2	4.111 1	0.000 0	0.000 0	1.250 0	0.000 0	1.000 0	0.000 0
72	0.750 0	1.000 0	4.444 4	0.000 0	0.000 0	2.000 0	0.000 0	0.250 0	0.000 0
73	0.756 9	3.222 2	0.000 0	2.500 0	0.250 0	0.000 0	2.000 0	0.000 0	0.000 0
74	0.187 5	3.666 7	0.750 0	0.000 0	0.750 0	3.000 0	0.000 0	0.000 0	0.000 0
75	1.666 7	0.000 0	2.750 0	3.500 0	0.000 0	0.000 0	0.000 0	0.000 0	1.000 0
76	0.020 4	2.982 8	0.000 0	0.000 0	6.498 3	0.000 0	0.000 0	0.000 0	0.000 0
77	0.305 6	2.350 8	2.040 0	0.000 0	3.611 1	1.423 6	0.000 0	0.000 0	0.000 0
78	0.178 9	3.288 3	1.312 5	0.000 0	2.673 6	2.361 1	0.000 0	0.000 0	0.000 0
79	0.152 8	3.427 2	1.222 2	0.000 0	2.534 7	2.500 0	0.000 0	0.000 0	0.000 0
80	0.870 0	2.631 9	0.000 0	3.062 5	2.250 0	0.000 0	1.361 1	0.000 0	0.000 0

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
81	0.614 7	1.958 3	3.923 6	0.000 0	1.000 0	1.611 1	0.000 0	1.000 0	0.000 0
82	0.566 1	2.236 1	3.645 8	0.000 0	0.250 0	3.111 1	0.000 0	0.250 0	0.000 0
83	0.660 0	1.444 4	4.250 0	0.000 0	1.000 0	2.500 0	0.000 0	0.111 1	0.000 0
84	0.637 2	3.708 3	0.000 0	2.361 1	1.361 1	0.000 0	2.250 0	0.000 0	0.000 0
85	0.498 3	2.847 2	3.222 2	0.000 0	0.111 1	2.500 0	0.000 0	1.000 0	0.000 0
86	0.142 5	3.944 4	0.611 1	0.000 0	1.972 2	3.250 0	0.000 0	0.000 0	0.000 0
87	1.381 9	1.583 3	2.000 0	3.361 1	0.000 0	1.000 0	0.250 0	0.000 0	1.000 0
88	1.375 0	1.250 0	2.333 3	3.222 2	0.000 0	1.000 0	1.000 0	0.000 0	0.250 0
89	1.291 7	1.722 2	2.611 1	2.750 0	0.000 0	0.250 0	1.000 0	0.000 0	1.000 0
90	1.194 4	0.000 0	6.944 4	0.000 0	0.000 0	0.000 0	0.000 0	2.250 0	0.000 0
91	0.562 5	2.666 7	3.222 2	0.000 0	0.250 0	2.500 0	0.000 0	1.000 0	0.000 0
92	0.520 8	4.416 7	0.000 0	1.750 0	0.750 0	0.000 0	3.000 0	0.000 0	0.000 0
93	0.015 6	3.023 6	0.000 0	0.000 0	7.989 7	0.000 0	0.000 0	0.000 0	0.000 0
94	0.290 8	2.446 4	2.027 8	0.000 0	5.034 7	1.463 6	0.000 0	0.000 0	0.000 0
95	0.159 3	3.406 4	1.290 0	0.000 0	4.074 7	2.423 6	0.000 0	0.000 0	0.000 0
96	0.122 9	3.593 9	1.173 6	0.000 0	3.887 2	2.611 1	0.000 0	0.000 0	0.000 0

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
97	0.833 3	2.814 4	0.000 0	3.040 0	3.611 1	0.000 0	1.423 6	0.000 0	0.000 0
98	0.567 8	2.211 9	3.852 5	0.000 0	2.250 0	1.784 7	0.000 0	1.000 0	0.000 0
99	0.493 1	2.628 6	3.484 4	0.000 0	1.361 1	3.423 6	0.000 0	0.250 0	0.000 0
100	0.496 7	2.586 9	3.477 5	0.000 0	1.361 1	3.562 5	0.000 0	0.111 1	0.000 0
101	0.611 1	1.694 4	4.160 0	0.000 0	2.250 0	2.722 2	0.000 0	0.062 5	0.000 0
102	0.580 0	3.961 9	0.000 0	2.312 5	2.673 6	0.000 0	2.361 1	0.000 0	0.000 0
103	0.415 0	3.288 3	3.034 7	0.000 0	1.173 6	2.861 1	0.000 0	1.000 0	0.000 0
104	0.392 5	3.427 2	2.847 2	0.000 0	0.562 5	4.222 2	0.000 0	0.250 0	0.000 0
105	0.527 8	4.149 4	0.000 0	2.222 2	2.534 7	0.000 0	2.500 0	0.000 0	0.000 0
106	0.118 1	4.090 6	0.562 5	0.000 0	3.347 2	3.361 1	0.000 0	0.000 0	0.000 0
107	0.107 8	4.177 2	0.472 2	0.000 0	3.256 9	3.500 0	0.000 0	0.000 0	0.000 0
108	1.265 8	2.131 9	1.861 1	3.312 5	1.000 0	1.250 0	0.361 1	0.000 0	1.000 0
109	1.168 6	2.548 6	1.583 3	3.173 6	0.250 0	2.000 0	1.111 1	0.000 0	0.250 0
110	1.240 0	1.805 6	2.187 5	3.125 0	1.000 0	1.250 0	1.250 0	0.000 0	0.111 1
111	1.149 4	2.319 4	2.562 5	2.611 1	1.000 0	0.361 1	1.250 0	0.000 0	1.000 0
112	0.961 9	1.486 1	6.006 9	0.000 0	0.000 0	1.361 1	0.000 0	2.250 0	0.000 0

续表

N ₀	x ₁	x ₂	x ₃	x ₄	x ₅	x ₆	x ₇	x ₈	x ₉
113	1.007 2	0.972 2	6.333 3	0.000 0	0.000 0	2.250 0	0.000 0	1.361 1	0.000 0
114	1.052 2	2.736 1	2.284 7	2.472 2	0.250 0	1.111 1	2.000 0	0.000 0	0.250 0
115	1.033 1	3.208 3	1.861 1	2.611 1	0.111 1	1.250 0	1.250 0	0.000 0	1.000 0
116	0.495 0	3.055 6	3.034 7	0.000 0	1.361 1	2.861 1	0.000 0	1.000 0	0.000 0
117	0.472 5	3.194 4	2.847 2	0.000 0	0.750 0	4.222 2	0.000 0	0.250 0	0.000 0
118	0.427 2	4.805 6	0.000 0	1.611 1	1.972 2	0.000 0	3.250 0	0.000 0	0.000 0
119	0.378 6	3.944 4	2.333 3	0.000 0	0.472 2	3.750 0	0.000 0	1.000 0	0.000 0
120	1.930 6	0.000 0	4.833 3	3.472 2	0.000 0	0.000 0	0.000 0	1.000 0	1.250 0
121	1.888 9	0.000 0	5.444 4	3.000 0	0.000 0	0.000 0	0.000 0	0.250 0	2.000 0
122	1.187 5	2.888 9	1.250 0	3.222 2	0.250 0	2.000 0	0.500 0	0.000 0	1.000 0
123	1.006 9	3.166 7	2.472 2	2.000 0	0.250 0	0.500 0	2.000 0	0.000 0	1.000 0
124	1.000 0	1.444 4	5.916 7	0.000 0	0.000 0	1.500 0	0.000 0	2.250 0	0.000 0
125	0.375 0	5.333 3	0.000 0	1.000 0	1.500 0	0.000 0	4.000 0	0.000 0	0.000 0
126	0.012 3	3.054 8	0.000 0	0.000 0	9.501 5	0.000 0	0.000 0	0.000 0	0.000 0
127	0.281 3	2.515 0	2.020 4	0.000 0	6.498 3	1.491 4	0.000 0	0.000 0	0.000 0
128	0.147 1	3.487 2	1.277 8	0.000 0	5.526 1	2.463 6	0.000 0	0.000 0	0.000 0
129	0.105 9	3.697 2	1.151 1	0.000 0	5.316 1	2.673 6	0.000 0	0.000 0	0.000 0
130	0.095 6	3.745 8	1.125 0	0.000 0	5.267 5	2.722 2	0.000 0	0.000 0	0.000 0

续表

N ₀	x ₁	x ₂	x ₃	x ₄	x ₅	x ₆	x ₇	x ₈	x ₉
131	0.315 9	2.646 9	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0
132	0.142 5	3.416 9	0.000 0	0.000 0	1.143 7	0.000 0	0.000 0	0.000 0	0.000 0
133	0.129 5	2.631 8	0.000 0	0.000 0	1.646 9	0.000 0	0.000 0	0.000 0	0.000 0
134	0.129 5	3.212 8	0.000 0	0.000 0	1.000 0	0.000 0	0.000 0	0.000 0	0.000 0
135	0.545 3	2.199 9	2.484 0	0.000 0	0.000 0	1.143 7	0.000 0	0.000 0	0.000 0
136	0.075 1	3.023 7	0.000 0	0.000 0	3.130 9	0.000 0	0.000 0	0.000 0	0.000 0
137	0.075 1	3.820 4	0.000 0	0.000 0	2.250 0	0.000 0	0.000 0	0.000 0	0.000 0
138	0.508 9	2.146 9	2.315 9	0.000 0	0.000 0	1.000 0	0.000 0	0.000 0	0.000 0
139	0.070 0	3.453 4	0.000 0	0.000 0	2.250 0	0.000 0	0.000 0	0.000 0	0.000 0
140	0.070 0	2.824 8	0.000 0	0.000 0	2.962 8	0.000 0	0.000 0	0.000 0	0.000 0
141	0.043 8	3.585 9	0.000 0	0.000 0	3.611 1	0.000 0	0.000 0	0.000 0	0.000 0
142	0.043 8	2.946 5	0.000 0	0.000 0	4.342 3	0.000 0	0.000 0	0.000 0	0.000 0
143	0.043 8	2.899 0	0.000 0	0.000 0	4.408 2	0.000 0	0.000 0	0.000 0	0.000 0
144	0.029 9	3.669 7	0.000 0	0.000 0	5.034 7	0.000 0	0.000 0	0.000 0	0.000 0
145	0.021 8	3.727 4	0.000 0	0.000 0	6.498 3	0.000 0	0.000 0	0.000 0	0.000 0
146	0.016 5	3.769 6	0.000 0	0.000 0	7.989 7	0.000 0	0.000 0	0.000 0	0.000 0
147	0.013 0	3.801 7	0.000 0	0.000 0	9.501 5	0.000 0	0.000 0	0.000 0	0.000 0

表 4-13A 106 个醇及其实验沸点 *b. p.* (单位: °C)

No	化合物	<i>b. p.</i>	No	化合物	<i>b. p.</i>
1	methanol	64.6	19	2 - methyl - 1 - pentanol	149.0
2	ethanol	78.0	20	2 - ethyl - 1 - butanol	147.0
3	propanol	97.1	21	2,3 - dimethyl - 1 - butanol	144.5
4	isopropanol	82.4	22	3,3 - dimethyl - 1 - butanol	143.0
5	butanol	117.6	23	2 - hexanol	140.0
6	2 - methyl - 1 - propanol	108.1	24	2,2 - dimethyl - 1 - butanol	136.5
7	2 - butanol	99.5	25	3 - hexanol	135.0
8	2 - methyl - 2 - propanol	82.4	26	3 - methyl - 2 - pentanol	134.3
9	pentanol	138.0	27	4 - methyl - 2 - pentanol	131.6
10	3 - methyl - 1 - butanol	131.0	28	2 - methyl - 3 - pentanol	126.5
11	2 - methyl - 1 - butanol	128.0	29	3 - methyl - 3 - pentanol	122.4
12	2 - pentanol	119.3	30	2 - methyl - 2 - pentanol	121.1
13	3 - pentanol	116.2	31	3,3 - dimethyl - 2 - butanol	120.4
14	3 - methyl - 2 - butanol	112.9	32	2,3 - dimethyl - 2 - butanol	118.4
15	2 - methyl - 2 - butanol	102.3	33	heptanol	176.4
16	hexanol	157.6	34	4 - methyl - 1 - hexanol	173.0
17	3 - methyl - 1 - pentanol	153.0	35	5 - methyl - 1 - hexanol	170.0
18	4 - methyl - 1 - pentanol	151.9	36	3 - methyl - 1 - hexanol	169.0

续表

No	化合物	b. p.	No	化合物	b. p.
37	2 - methyl - 1 - hexanol	164.0	55	4 - methyl - 1 - heptanol	188.0
38	2,4 - dimethyl - 1 - pentanol	159.0	56	6 - methyl - 1 - heptanol	188.6
39	2 - heptanol	159.0	57	2,5 - dimethyl - 1 - hexanol	179.5
40	2 - methyl - 2 - hexanol	143.0	58	2,4,4 - trimethyl - 1 - pentanol	168.5
41	5 - methyl - 2 - hexanol	151.0	59	2 - octanol	180.0
42	2,3 - dimethyl - 2 - pentanol	139.7	60	2 - methyl - 2 - heptanol	156.0
43	2,4 - dimethyl - 2 - pentanol	133.1	61	3 - methyl - 2 - heptanol	166.1
44	2,3,3 - trimethyl - 2 - butanol	131.0	62	5 - methyl - 2 - heptanol	170.0
45	3 - heptanol	157.0	63	2,3 - dimethyl - 2 - hexanol	160.0
46	2 - methyl - 3 - hexanol	143.0	64	2,5 - dimethyl - 2 - hexanol	154.5
47	3 - methyl - 3 - hexanol	143.0	65	3,4 - dimethyl - 2 - hexanol	165.5
48	5 - methyl - 3 - hexanol	148.0	66	2 - methyl - 3 - ethyl - 2 - pentanol	156.0
49	3 - ethyl - 3 - pentanol	142.0	67	2,3,4 - trimethyl - 2 - pentanol	147.5
50	2,2 - dimethyl - 3 - pentanol	135.0	68	3 - octanol	171.0
51	2,3 - dimethyl - 3 - pentanol	139.7	69	2 - methyl - 3 - heptanol	167.5
52	2,4 - dimethyl - 3 - pentanol	138.7	70	3 - methyl - 3 - heptanol	163.0
53	4 - heptanol	156.0	71	4 - methyl - 3 - heptanol	170.0
54	octanol	195.3	72	5 - methyl - 3 - heptanol	172.0

续表

No	化合物	<i>b. p.</i>	No	化合物	<i>b. p.</i>
73	6 - methyl - 3 - heptanol	174.0	90	2,2,3 - trimethyl - 3 - hexanol	156.0
74	2 - methyl - 3 - ethyl - 3 - pentanol	160.0	91	4 - nonanol	192.5
75	2,2,4 - trimethyl - 3 - pentanol	150.5	92	2,6 - dimethyl - 4 - heptanol	174.5
76	2,3,4 - trimethyl - 3 - pentanol	156.5	93	3,5 - dimethyl - 4 - heptanol	187.0
77	4 - octanol	176.3	94	4 - ethyl - 4 - heptanol	182.0
78	2 - methyl - 4 - heptanol	164.0	95	5 - nonanol	193.0
79	3 - methyl - 4 - heptanol	162.0	96	7 - methyl - 1 - octanol	206.0
80	4 - methyl - 4 - heptanol	161.0	97	decanol	231.1
81	nonanol	213.3	98	3,7 - dimethyl - 1 - octanol	212.5
82	2,2 - diethyl - 1 - pentanol	192.0	99	2 - decanol	211.0
83	3,5,5 - trimethyl - 1 - hexanol	193.0	100	2,7 - dimethyl - 3 - octanol	193.5
84	2 - nonanol	193.5	101	3,6 - dimethyl - 3 - octanol	202.2
85	2 - methyl - 2 - octanol	178.0	102	3 - ethyl - 3 - octanol	199.0
86	2,6 - dimethyl - 2 - heptanol	173.0	103	3 - ethyl - 2 - methyl - 3 - heptanol	193.0
87	3 - nonanol	195.0	104	4 - decanol	210.5
88	2,6 - dimethyl - 3 - heptanol	175.0	105	2,6 - dimethyl - 4 - octanol	195.0
89	3,6 - dimethyl - 3 - heptanol	173.0	106	undecanol	245.0

表 4-13B 106 个醇的 MEDV-4 矢量描述子

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	1.564 5	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0
2	0.362 7	2.564 5	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0
3	0.157 3	3.177 2	0.000 0	0.000 0	1.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0
4	0.975 4	0.000 0	3.564 5	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0
5	0.087 4	3.445 6	0.000 0	0.000 0	2.250 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0
6	0.564 6	2.064 5	2.362 7	0.000 0	0.000 0	1.000 0	0.000 0	0.000 0	0.000 0	0.000 0
7	0.631 1	1.612 7	2.814 5	0.000 0	0.000 0	1.000 0	0.000 0	0.000 0	0.000 0	0.000 0
8	1.838 1	0.000 0	0.000 0	4.564 5	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0
9	0.055 5	3.595 6	0.000 0	0.000 0	3.611 1	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0
10	0.424 8	2.649 5	2.157 3	0.000 0	1.000 0	1.250 0	0.000 0	0.000 0	0.000 0	0.000 0
11	0.355 8	3.332 9	1.612 7	0.000 0	0.250 0	2.000 0	0.000 0	0.000 0	0.000 0	0.000 0
12	0.512 6	2.131 1	2.675 7	0.000 0	1.000 0	1.250 0	0.000 0	0.000 0	0.000 0	0.000 0
13	0.377 1	2.947 6	2.064 5	0.000 0	0.250 0	2.000 0	0.000 0	0.000 0	0.000 0	0.000 0
14	1.149 5	0.000 0	5.677 2	0.000 0	0.000 0	0.000 0	0.000 0	1.000 0	0.000 0	0.000 0
15	1.354 9	1.862 7	0.000 0	3.814 5	0.000 0	0.000 0	1.000 0	0.000 0	0.000 0	0.000 0
16	0.038 4	3.691 1	0.000 0	0.000 0	5.034 7	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
17	0.254 1	3.799 4	1.407 3	0.000 0	1.361 1	2.250 0	0.000 0	0.000 0	0.000 0	0.000 0
18	0.361 1	2.931 8	2.087 4	0.000 0	2.250 0	1.361 1	0.000 0	0.000 0	0.000 0	0.000 0
19	0.275 3	3.732 9	1.473 8	0.000 0	1.361 1	2.250 0	0.000 0	0.000 0	0.000 0	0.000 0
20	0.237 3	4.323 6	0.862 7	0.000 0	0.750 0	3.000 0	0.000 0	0.000 0	0.000 0	0.000 0
21	0.804 3	2.036 8	4.270 0	0.000 0	0.000 0	1.250 0	0.000 0	1.000 0	0.000 0	0.000 0
22	1.012 2	3.010 6	0.000 0	3.157 3	1.000 0	0.000 0	1.250 0	0.000 0	0.000 0	0.000 0
23	0.458 2	2.392 1	2.627 0	0.000 0	2.250 0	1.361 1	0.000 0	0.000 0	0.000 0	0.000 0
24	0.874 2	3.832 9	0.000 0	2.612 7	0.250 0	0.000 0	2.000 0	0.000 0	0.000 0	0.000 0
25	0.284 7	3.368 8	1.925 7	0.000 0	1.361 1	2.250 0	0.000 0	0.000 0	0.000 0	0.000 0
26	0.892 1	1.518 4	4.788 4	0.000 0	0.000 0	1.250 0	0.000 0	1.000 0	0.000 0	0.000 0
27	0.912 5	1.112 7	5.055 2	0.000 0	0.000 0	2.000 0	0.000 0	0.250 0	0.000 0	0.000 0
28	0.846 9	1.584 9	4.788 4	0.000 0	0.000 0	1.250 0	0.000 0	1.000 0	0.000 0	0.000 0
29	0.962 0	3.447 6	0.000 0	3.064 5	0.250 0	0.000 0	2.000 0	0.000 0	0.000 0	0.000 0
30	1.187 8	2.492 2	0.000 0	3.675 7	1.000 0	0.000 0	1.250 0	0.000 0	0.000 0	0.000 0
31	1.917 9	0.000 0	3.314 5	3.612 7	0.000 0	0.000 0	0.000 0	0.000 0	1.000 0	0.000 0
32	1.984 4	0.000 0	2.862 7	4.064 5	0.000 0	0.000 0	0.000 0	0.000 0	1.000 0	0.000 0
33	0.028 1	3.757 2	0.000 0	0.000 0	6.498 3	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0
34	0.205 0	4.027 3	1.337 4	0.000 0	2.673 6	2.361 1	0.000 0	0.000 0	0.000 0	0.000 0

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
35	0.326 8	3.099 2	2.055 5	0.000 0	3.611 1	1.423 6	0.000 0	0.000 0	0.000 0	0.000 0
36	0.188 3	4.144 9	1.268 4	0.000 0	2.534 7	2.500 0	0.000 0	0.000 0	0.000 0	0.000 0
37	0.235 7	3.939 5	1.425 2	0.000 0	2.673 6	2.361 1	0.000 0	0.000 0	0.000 0	0.000 0
38	0.643 4	2.846 8	3.783 4	0.000 0	0.250 0	3.111 1	0.000 0	0.250 0	0.000 0	0.000 0
39	0.428 9	2.550 2	2.604 5	0.000 0	3.611 1	1.423 6	0.000 0	0.000 0	0.000 0	0.000 0
40	1.110 9	2.815 7	0.000 0	3.627 0	2.250 0	0.000 0	1.361 1	0.000 0	0.000 0	0.000 0
41	0.803 8	1.603 3	4.839 5	0.000 0	1.000 0	2.500 0	0.000 0	0.111 1	0.000 0	0.000 0
42	1.678 4	1.629 5	2.112 7	3.925 7	0.000 0	1.000 0	0.250 0	0.000 0	1.000 0	0.000 0
43	1.650 2	1.362 7	2.379 5	3.786 8	0.000 0	1.000 0	1.000 0	0.000 0	0.250 0	0.000 0
44	2.863 9	0.000 0	0.000 0	8.177 2	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	1.000 0
45	0.240 6	3.584 8	1.877 0	0.000 0	2.673 6	2.361 1	0.000 0	0.000 0	0.000 0	0.000 0
46	0.732 0	2.117 2	4.600 9	0.000 0	1.000 0	1.611 1	0.000 0	1.000 0	0.000 0	0.000 0
47	0.821 0	3.979 9	0.000 0	2.925 7	1.361 1	0.000 0	2.250 0	0.000 0	0.000 0	0.000 0
48	0.662 1	2.461 5	4.256 6	0.000 0	0.250 0	3.111 1	0.000 0	0.250 0	0.000 0	0.000 0
49	0.659 4	4.754 8	0.000 0	2.314 5	0.750 0	0.000 0	3.000 0	0.000 0	0.000 0	0.000 0
50	1.566 7	1.696 0	2.564 5	3.473 8	0.000 0	1.000 0	0.250 0	0.000 0	1.000 0	0.000 0
51	1.542 9	1.834 9	2.723 8	3.314 5	0.000 0	0.250 0	1.000 0	0.000 0	1.000 0	0.000 0
52	1.379 2	0.000 0	7.734 4	0.000 0	0.000 0	0.000 0	0.000 0	2.250 0	0.000 0	0.000 0

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
53	0.202 6	3.745 0	1.786 8	0.000 0	2.534 7	2.500 0	0.000 0	0.000 0	0.000 0	0.000 0
54	0.021 5	3.805 8	0.000 0	0.000 0	7.989 7	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0
55	0.146 1	4.343 5	1.198 5	0.000 0	3.887 2	2.611 1	0.000 0	0.000 0	0.000 0	0.000 0
56	0.306 2	3.210 3	2.038 4	0.000 0	5.034 7	1.463 6	0.000 0	0.000 0	0.000 0	0.000 0
57	0.564 1	3.222 6	3.605 7	0.000 0	1.361 1	3.562 5	0.000 0	0.111 1	0.000 0	0.000 0
58	1.261 4	3.159 3	1.696 0	3.198 5	0.250 0	2.000 0	1.111 1	0.000 0	0.250 0	0.000 0
59	0.411 2	2.656 3	2.592 3	0.000 0	5.034 7	1.463 6	0.000 0	0.000 0	0.000 0	0.000 0
60	1.069 3	3.013 8	0.000 0	3.604 5	3.611 1	0.000 0	1.423 6	0.000 0	0.000 0	0.000 0
61	0.737 3	2.298 6	4.529 7	0.000 0	2.250 0	1.784 7	0.000 0	1.000 0	0.000 0	0.000 0
62	0.635 5	2.761 4	4.067 0	0.000 0	1.361 1	3.562 5	0.000 0	0.111 1	0.000 0	0.000 0
63	1.552 9	2.203 0	1.973 8	3.877 0	1.000 0	1.250 0	0.361 1	0.000 0	1.000 0	0.000 0
64	1.496 5	1.964 4	2.212 4	3.689 5	1.000 0	1.250 0	1.250 0	0.000 0	0.111 1	0.000 0
65	1.161 3	1.511 0	6.730 4	0.000 0	0.000 0	1.361 1	0.000 0	2.250 0	0.000 0	0.000 0
66	1.462 7	2.981 3	1.362 7	3.786 8	0.250 0	2.000 0	0.500 0	0.000 0	1.000 0	0.000 0
67	2.252 0	0.000 0	4.992 2	4.036 8	0.000 0	0.000 0	0.000 0	1.000 0	1.250 0	0.000 0
68	0.216 1	3.718 4	1.854 5	0.000 0	4.074 7	2.423 6	0.000 0	0.000 0	0.000 0	0.000 0
69	0.675 7	2.395 7	4.529 7	0.000 0	2.250 0	1.784 7	0.000 0	1.000 0	0.000 0	0.000 0
70	0.754 4	4.258 4	0.000 0	2.877 0	2.673 6	0.000 0	2.361 1	0.000 0	0.000 0	0.000 0

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
71	0.522 9	3.472 1	3.712 0	0.000 0	1.173 6	2.861 1	0.000 0	1.000 0	0.000 0	0.000 0
72	0.479 1	3.677 5	3.457 9	0.000 0	0.562 5	4.222 2	0.000 0	0.250 0	0.000 0	0.000 0
73	0.573 9	2.858 5	4.067 0	0.000 0	1.361 1	3.562 5	0.000 0	0.111 1	0.000 0	0.000 0
74	1.191 7	3.392 1	2.584 9	2.564 5	0.250 0	0.500 0	2.000 0	0.000 0	1.000 0	0.000 0
75	2.161 5	0.000 0	5.510 6	3.584 9	0.000 0	0.000 0	0.000 0	1.000 0	1.250 0	0.000 0
76	2.186 3	0.000 0	5.669 8	3.564 5	0.000 0	0.000 0	0.000 0	0.250 0	2.000 0	0.000 0
77	0.163 4	3.936 6	1.738 2	0.000 0	3.887 2	2.611 1	0.000 0	0.000 0	0.000 0	0.000 0
78	0.567 8	2.900 2	4.095 2	0.000 0	1.361 1	3.423 6	0.000 0	0.250 0	0.000 0	0.000 0
79	0.511 0	3.493 4	3.712 0	0.000 0	1.173 6	2.861 1	0.000 0	1.000 0	0.000 0	0.000 0
80	0.690 3	4.467 2	0.000 0	2.786 8	2.534 7	0.000 0	2.500 0	0.000 0	0.000 0	0.000 0
81	0.016 9	3.842 8	0.000 0	0.000 0	9.501 5	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0
82	0.372 9	6.3530	0.000 0	0.973 8	2.833 3	0.000 0	4.250 0	0.000 0	0.000 0	0.000 0
83	1.140 1	3.683 3	1.490 6	3.166 6	1.361 1	2.250 0	1.173 6	0.000 0	0.250 0	0.000 0
84	0.399 8	2.732 6	2.585 0	0.000 0	6.498 3	1.491 4	0.000 0	0.000 0	0.000 0	0.000 0
85	1.044 3	3.147 7	0.000 0	3.592 3	5.034 7	0.000 0	1.463 6	0.000 0	0.000 0	0.000 0
86	1.413 3	2.301 8	2.135 5	3.644 5	2.250 0	1.361 1	1.361 1	0.000 0	0.062 5	0.000 0
87	0.201 0	3.809 8	1.842 3	0.000 0	5.526 1	2.463 6	0.000 0	0.000 0	0.000 0	0.000 0
88	1.036 8	1.589 4	6.799 7	0.000 0	1.000 0	2.861 1	0.000 0	1.173 6	0.000 0	0.000 0

续表

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
89	1.127 7	3.469 6	2.189 9	2.939 5	1.361 1	1.312 5	2.250 0	0.000 0	0.111 1	0.000 0
90	2.187 8	2.651 9	0.000 0	7.100 9	1.000 0	0.000 0	1.611 1	0.000 0	0.000 0	1.000 0
91	0.141 4	4.055 4	1.715 7	0.000 0	5.316 1	2.673 6	0.000 0	0.000 0	0.000 0	0.000 0
92	0.960 7	1.975 4	6.483 6	0.000 0	0.250 0	4.222 2	0.000 0	0.562 5	0.000 0	0.000 0
93	0.881 9	3.019 6	5.859 4	0.000 0	0.062 5	2.722 2	0.000 0	2.250 0	0.000 0	0.000 0
94	0.439 9	5.579 9	0.000 0	2.036 8	3.256 9	0.000 0	3.500 0	0.000 0	0.000 0	0.000 0
95	0.126 7	4.113 4	1.689 5	0.000 0	5.267 5	2.722 2	0.000 0	0.000 0	0.000 0	0.000 0
96	0.292 9	3.289 5	2.028 1	0.000 0	6.498 3	1.491 4	0.000 0	0.000 0	0.000 0	0.000 0
97	0.013 7	3.872 1	0.000 0	0.000 0	11.028 9	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0
98	0.435 9	3.836 9	3.305 4	0.000 0	3.887 2	4.040 0	0.000 0	0.062 5	0.000 0	0.000 0
99	0.392 0	2.790 1	2.580 2	0.000 0	7.989 7	1.511 8	0.000 0	0.000 0	0.000 0	0.000 0
100	0.973 0	1.881 8	6.673 9	0.000 0	2.250 0	3.145 8	0.000 0	1.102 5	0.000 0	0.000 0
101	0.952 1	4.643 2	1.439 9	2.917 0	1.762 2	2.312 5	2.312 5	0.000 0	0.111 1	0.000 0
102	0.456 3	5.572 5	0.000 0	2.104 5	4.788 3	0.000 0	3.423 6	0.000 0	0.000 0	0.000 0
103	0.985 7	4.234 3	2.513 8	2.377 0	2.673 6	0.673 6	2.361 1	0.000 0	1.000 0	0.000 0
104	0.127 9	4.137 3	1.703 4	0.000 0	6.787 9	2.713 6	0.000 0	0.000 0	0.000 0	0.000 0
105	0.775 2	3.207 0	5.672 7	0.000 0	0.562 5	5.373 3	0.000 0	0.562 5	0.000 0	0.000 0
106	0.011 3	3.895 8	0.000 0	0.000 0	12.568 7	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0

我们曾建立了 10 参数方程^[28]，并对其可估计与可预测能力进行了分析，表明模型具有很好的估计与预测能力。这个模型如下：

$$\begin{aligned}
 b.p. = & -(130.249\ 2 \pm 12.908\ 8) + (126.796\ 0 \pm 8.600\ 0)x_1 + \\
 & (62.598\ 6 \pm 3.632\ 5)x_2 + (23.198\ 3 \pm 1.272\ 6)x_3 - \\
 & (5.572\ 7 \pm 1.050\ 0)x_4 + (10.455\ 3 \pm 0.219\ 5)x_5 - \\
 & (16.377\ 1 \pm 1.905\ 1)x_6 - (34.303\ 1 \pm 2.849\ 8)x_7 - \\
 & (37.175\ 7 \pm 3.480\ 7)x_8 - (48.167\ 6 \pm 4.395\ 7)x_9 - \\
 & (64.546\ 1 \pm 6.100\ 8)x_{10} \quad (4-21)
 \end{aligned}$$

$n = 106, m = 10, r = 0.995\ 1, RMSE = 3.23\ K, F = 964.17$ (建模时)
 $q = 0.989\ 9, RMSV = 4.65\ K$ (LOO 校验)

上述结果表明，醇沸点与 MEDV - 4 有良好相关关系。但是从各个样本的 LOO 法预测误差看，甲醇 (No. 1)、2,3,3 - 三甲基 - 2 - 丁醇 (No. 44) 和 2,2,3 - 三甲基 - 3 - 己醇 (No. 90) 三个化合物误差比其他所有化合物都大，分别达 21.7、-17.4 和 17.4 K，这可从它们的 MEDV 矢量描述子取非零值的特征得以解释。它们是离群值，可以预期，从数据集中删除这三个化合物，模型质量应当更高。由此思路建立不含 x_{10} 的 9 参数 QSPR 模型如下：

$$\begin{aligned}
 b.p. = & (139.373\ 8 \pm 11.940\ 5) + (137.421\ 8 \pm 9.255\ 6)x_1 + \\
 & (63.109\ 8 \pm 3.352\ 1)x_2 + (21.708\ 5 \pm 1.455\ 8)x_3 - \\
 & (8.850\ 3 \pm 1.808\ 5)x_4 + (10.645\ 2 \pm 0.231\ 5)x_5 - \\
 & (15.945\ 6 \pm 1.797\ 9)x_6 - (33.589\ 6 \pm 2.661\ 6)x_7 - \\
 & (37.017\ 1 \pm 3.229\ 6)x_8 - (47.977\ 1 \pm 4.056\ 3)x_9 \quad (4-22)
 \end{aligned}$$

$n = 103, m = 9, r = 0.995\ 6, RMSE = 2.99\ K, F = 1\ 156.91$ (建模时)
 $q = 0.993\ 9, RMSV = 3.51\ K$ (LOO 校验)

结果如所预料，质量大大提高。由 (4-22) 式估计的醇沸点 (BP_{EST}) 及实验沸点 BP_{OBS} 相关情况见图 4-11。

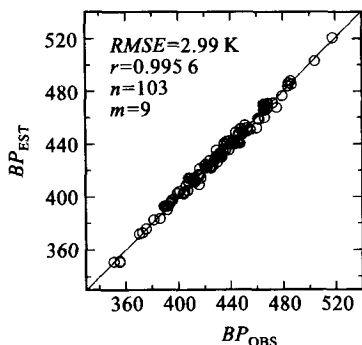


图 4-11 103 个醇估计与实验
沸点相关图

4.4 芳烃化合物结构表征与性质相关

前节研究了部分不饱和烃及含杂原子化合物的表征与性质相关。本节将 MEDV-4 进一步扩展到更加复杂的取代芳烃及多环芳烃与含杂原子多环芳烃体系，以评价 MEDV 在这些体系的可应用性。

4.4.1 66 个不含杂原子芳烃与沸点

从 Basak 等人的报告^[7]中选择 66 个单取代苯和多环芳烃（简称 PAH）及其相应实验沸点为研究对象。各芳烃分子结构仍以 MEDV 矢量描述子表征。由于这些芳烃分子中均不含第 4 原子类型原子，因此它们的 MEDV 中实际上只有 6 个描述子具有非零样本值。另 4 个描述子即 x_4, x_7, x_9, x_{10} 均为零。这些化合物的编号、名称、6 个 MEDV 描述子 (x) 及实验沸点值 (BP_{OBS}) 一并列入表 4-14。由表可知，64 号与 65 号两个多环芳烃的 6 个描述子是相同的，说明 MEDV 不能分辨这两个化合物。然而，这两个化合

表 4-14 66 个芳烃的 6 个 MEDV-4 描述子及其实验沸点 $b.p.$ (单位:K)

No	化合物	x_1	x_2	x_3	x_5	x_6	x_8	$b.p.$
1	benzene	0	0	0	8.959 1	0	0	353.3
2	toluene	0	0.846 0	1.000 0	5.972 7	2.986 4	0	383.8
3	ethylbenzene	0	1.409 0	0.250 0	6.818 7	3.986 4	0	409.4
4	<i>o</i> -xylene	0.116 1	1.157 8	2.534 1	4.130 1	3.685 3	1.143 7	417.6
5	<i>m</i> -xylene	0.066 8	1.449 2	2.242 8	3.272 3	5.400 9	0.285 9	412.3
6	<i>p</i> -xylene	0.043 3	1.553 8	2.138 1	3.113 4	5.718 6	0.127 1	411.6
7	<i>n</i> -propylbenzene	0	1.493 2	0.111 1	8.227 7	4.236 4	0	432.4
8	1-methyl-2-ethylbenzene	0.064 6	1.987 9	1.633 1	4.709 0	4.952 4	1.143 7	438.4
9	1-methyl-3-ethylbenzene	0.042 2	2.133 6	1.438 2	3.996 9	6.522 3	0.285 9	434.5
10	1-methyl-4-ethylbenzene	0.029 7	2.185 9	1.362 4	3.890 3	6.787 6	0.127 1	435.2
11	1,2,3-trimethylbenzene	0.298 9	1.226 9	4.311 0	2.573 4	3.812 4	2.573 4	449.3
12	1,2,4-trimethylbenzene	0.226 2	1.622 9	3.915 0	1.556 7	5.845 6	1.556 7	442.6
13	1,3,5-trimethylbenzene	0.200 3	1.809 5	3.728 4	0.857 8	7.243 5	0.857 8	437.9
14	<i>n</i> -butylbenzene	0	1.522 9	0.062 5	9.720 8	4.347 5	0	456.5
15	1,2-diethylbenzene	0.041 1	2.715 0	0.732 2	5.404 0	6.219 4	1.143 7	456.6
16	1,3-diethylbenzene	0.029 0	2.768 8	0.633 5	4.788 2	7.643 6	0.285 9	454.3

续表

No	化合物	x_1	x_2	x_3	x_5	x_6	x_8	$b, p.$
17	1,4 - diethylbenzene	0.021 6	2.790 7	0.586 6	4.710 6	7.856 7	0.127 1	457.0
18	1 - methyl - 2 - propylbenzene	0.041 1	2.188 1	1.442 8	6.001 9	5.318 4	1.143 7	458.0
19	1 - methyl - 3 - propylbenzene	0.029 0	2.284 5	1.274 7	5.339 1	6.839 0	0.285 9	455.0
20	1 - methyl - 4 - propylbenzene	0.021 6	2.313 4	1.209 8	5.256 0	7.080 9	0.127 1	457.0
21	1,2 - dimethyl - 3 - ethylbenzene	0.222 8	2.178 4	3.355 4	3.030 9	5.200 8	2.573 4	467.1
22	1,2 - dimethyl - 4 - ethylbenzene	0.187 9	2.376 4	3.084 7	2.212 2	7.036 1	1.556 7	463.0
23	1,3 - dimethyl - 2 - ethylbenzene	0.195 9	2.324 0	3.259 1	2.885 2	5.346 5	2.573 4	463.2
24	1,3 - dimethyl - 4 - ethylbenzene	0.161 0	2.522 0	2.988 3	2.066 6	7.181 8	1.556 7	461.6
25	1,3 - dimethyl - 5 - ethylbenzene	0.151 1	2.615 4	2.869 1	1.461 0	8.486 3	0.857 8	456.6
26	1,4 - dimethyl - 2 - ethylbenzene	0.150 1	2.574 4	2.959 4	2.014 2	7.234 1	1.556 7	459.3
27	1,2,3,4 - tetramethylbenzene	0.525 1	1.157 8	6.226 1	1.143 7	3.685 3	4.130 1	478.2
28	1,2,3,5 - tetramethylbenzene	0.475 8	1.449 2	5.934 7	0.285 9	5.400 9	3.272 3	471.4
29	1,2,4,5 - tetramethylbenzene	0.452 3	1.553 8	5.830 1	0.127 1	5.718 6	3.113 4	470.0
30	naphthalene	0	0	0	10.114 9	7.370 6	1.143 7	491.2
31	acenaphthylene	0	0	0	8.717 0	11.961 3	4.288 9	543.2
32	acenaphthene	0	0	0	8.483 1	11.552 9	4.288 9	552.2
33	fluorene	0	0	0	10.307 0	11.890 7	4.253 0	567.2

续表

No	化合物	x_1	x_2	x_3	x_5	x_6	x_8	$b, p.$
34	phenanthrene	0	0	0	12.096 5	12.910 6	4.130 1	611.2
35	anthracene	0	0	0	11.488 2	14.352 3	3.113 4	613.2
36	4H - cyclopenta[def]phenanthrene	0	0	0	9.185 7	16.021 4	8.065 4	632.2
37	fluoranthene	0	0	0	11.525 4	15.325 1	9.117 9	656.2
38	pyrene	0	0	0	10.917 2	17.195 7	7.942 4	666.2
39	benzo[a]fluorene	0	0	0	12.363 0	17.267 3	7.620 8	676.2
40	benzo[b]fluorene	0	0	0	11.727 7	18.786 3	6.559 5	671.2
41	benzo[c]fluorene	0	0	0	12.422 5	17.103 3	7.850 9	679.2
42	benzo[ghi]fluoranthene	0	0	0	10.471 0	19.173 7	13.383 0	695.2
43	cyclopenta[cd]pyrene	0	0	0	9.711 8	21.090 0	12.096 3	712.2
44	chrysene	0	0	0	14.137 6	18.473 8	7.432 2	704.2
45	benz[a]anthracene	0	0	0	13.529 4	19.915 5	6.415 6	698.2
46	triphenylene	0	0	0	14.904 0	16.620 0	8.959 1	702.2
47	naphthacene	0	0	0	12.904 6	21.395 6	5.317 6	713.2
48	benzo[b]fluoranthene	0	0	0	13.679 9	20.510 6	13.057 4	754.2
49	benzo[j]fluoranthene	0	0	0	13.679 9	20.513 9	13.031 6	753.2
50	benzo[k]fluoranthene	0	0	0	12.981 8	22.215 0	11.719 2	754.2

续表

No	化合物	x_1	x_2	x_3	x_5	x_6	x_8	$b \cdot p$
51	benzo[a]pyrene	0	0	0	13.051 7	22.462 6	11.774 8	769.2
52	benzo[e]pyrene	0	0	0	13.763 1	20.734 9	13.168 6	766.2
53	perylene	0	0	0	13.738 8	20.818 1	13.087 2	770.2
54	anthanthrene	0	0	0	12.024 8	26.209 4	16.433 2	820.2
55	benzo[ghi]perylene	0	0	0	12.681 1	24.607 6	17.693 9	815.2
56	indeno[1,2,3-cd]fluoranthene	0	0	0	13.117 0	22.752 2	18.355 3	804.2
57	indeno[1,2,3-cd]pyrene	0	0	0	12.590 7	24.421 7	17.501 3	807.2
58	dibenz[a,c]anthracene	0	0	0	16.396 4	23.648 1	11.560 4	808.2
59	dibenz[a,h]anthracene	0	0	0	15.598 8	25.552 5	9.850 4	808.2
60	dibenz[a,j]anthracene	0	0	0	15.607 6	25.547 1	9.912 5	804.2
61	picene	0	0	0	16.207 1	24.110 8	10.867 0	792.2
62	coronene	0	0	0	11.658 1	28.238 2	22.534 9	863.2
63	dibenz[a,e]pyrene	0	0	0	15.957 1	26.024 9	17.316 7	865.2
64	dibenz[a,h]pyrene	0	0	0	15.214 6	27.803 3	15.739 8	869.2
65	dibenz[a,i]pyrene	0	0	0	15.214 6	27.803 3	15.739 8	867.2
66	dibenz[a,l]pyrene	0	0	0	16.044 4	25.799 7	17.612 5	868.2

物结构相似，沸点也几近相同 (869.2 K 和 867.2 K) 故不会明显影响模型质量。从 MEDV 描述子之间的相关分析来看，描述子 x_1 与描述子 x_3 之间的相关系数达到 0.961 6，说明它们高度相关，建立模型时可考虑从中删除一个描述子，也许能获得更加稳定的线性模型。通过 MLR 方法建立 6 参数模型如下：

$$\begin{aligned}
 b.p. = & (286.159\ 8 \pm 7.671\ 9) + (169.536\ 9 \pm 43.434\ 2)x_1 + \\
 & (7.996\ 4 \pm 2.052\ 8)x_2 + (2.234\ 2 \pm 3.891\ 4)x_3 + \\
 & (10.819\ 4 \pm 0.790\ 0)x_5 + (12.027\ 7 \pm 0.431\ 7)x_6 + \\
 & (5.471\ 2 \pm 0.464\ 5)x_8 \quad (4-23)
 \end{aligned}$$

$n=66, m=6, r^2=0.997\ 6, r=0.998\ 8, RMSE=7.91\ K, F=4\ 049.56$
 (建模时)
 $q^2=0.996\ 7, q=0.998\ 4, RMSV=9.19\ K$ (LOO 校验)

由 (4-23) 式可知，描述子 x_1 与描述子 x_3 的回归系数相对标准偏差较大，需要加以改进。虽然描述子 x_3 有 28 个非零值，大于描述子 x_1 的 24 个非零样本值，但是回归系数则是前者小于后者很多，故删除描述子 x_3 重建 5 参数 QSPR 模型：

$$\begin{aligned}
 b.p. = & (288.419\ 3 \pm 6.548\ 8) + (193.110\ 6 \pm 14.088\ 5)x_1 + \\
 & (8.214\ 6 \pm 2.006\ 1)x_2 + (10.591\ 5 \pm 0.679\ 3)x_5 + \\
 & (12.076\ 9 \pm 0.420\ 7)x_6 + (5.445\ 5 \pm 0.459\ 7)x_8 \quad (4-24)
 \end{aligned}$$

$n=66, m=5, r=0.9988, RMSE=7.93\ K, F=4\ 914.31$ (建模时)
 $q=0.998\ 4, RMSV=9.14\ K$ (LOO 校验)

很明显，(4-24) 式中各个回归系数的标准偏差均有降低，特别是描述子 x_1 的回归系数标准偏差降低非常明显。此外，虽然均方根误差没有显著性差异，但 F 统计量增加却是显著的，说明该模型具有更重要的统计意义。LOO 分析预测也表明，5 参数模型更优。由 (4-24) 式估计的沸点 (BP_{66}) 与实验沸点 (BP_{OBS}) 的

相关情况见图 4-12A。

由图 4-12A 可知，第一个样本误差明显偏大。由 LOO 分析预测该样本的误差达 40.1K，是典型的离群值。这是由于该样本不同于所有其他样本，它的 MEDV 描述子中，只有描述子 x_5 为非零值 其他 5 个 MEDV 描述子均为零。除去该样本后，重建 65 个样本沸点与 MEDV 的相关模型如下：

$$\begin{aligned}
 b.p. = & (295.134\ 5 \pm 5.731\ 9) + (188.961\ 0 \pm 12.004\ 1)x_1 + \\
 & (6.579\ 4 \pm 1.737\ 3)x_2 + (10.963\ 0 \pm 0.582\ 3)x_3 + \\
 & (11.326\ 8 \pm 0.388\ 9)x_6 + (5.828\ 1 \pm 0.398\ 4)x_8 \quad (4-25)
 \end{aligned}$$

$n=65, m=5, r^2=0.998\ 2, r=0.999\ 1, RMSE=6.73\ K, F=6\ 554.57$
 (建模时)
 $q^2=0.997\ 8, q=0.998\ 9, RMSV=7.44\ K$ (LOO 校验)

与 (4-24) 式结果比较，5 参数模型结果的 RMSE 和 F 统计量都有改善，由 4-25 式估计的沸点 BP_{65} 与实验观测 BP_{OBS} 的相关情况如图 4-12B 所示。

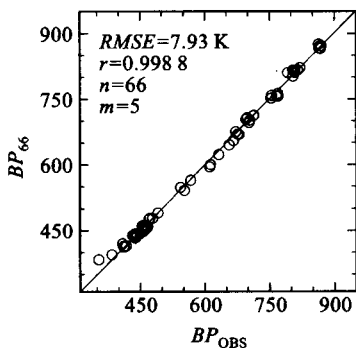


图 4-12A 66 个芳烃估计与实验沸点相关图

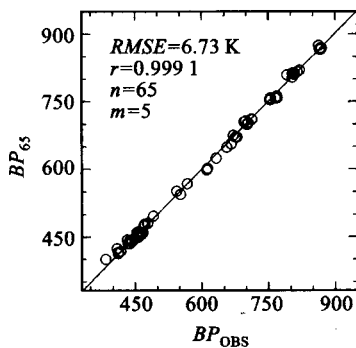


图 4-12B 65 个芳烃估计与实验沸点相关图

4.4.2 209个多环芳烃与色谱保留指数

选择 Lee 的报告^[29]中 209 个多环芳烃 (polycycle aromatic hydrocarbon 简称 PAH) 及其气相色谱保留指数为研究对象。由于这些 PAH 均不含第 4 原子类型原子, 故实际上只有 6 个 MEDV 描述子具有非零样本值, 其他描述子 x_4, x_7, x_9, x_{10} 全部为零。这 6 个 MEDV 描述子 x_1, x_2, x_3, x_5, x_6 和 x_8 的非零样本数分别为 35, 102, 102, 209, 209 和 209, 显然都大于样本数的 10%。表 4-15 列出了 209 个芳烃的气相色谱保留指数实验值与 6 个非零 MEDV 描述子值^[30]。

对 209 个 PAH 样本的 MEDV 矢量, 计算两两间的模式距离, 结果发现有 7 对化合物即 36 号与 37 号、55 号和 56 号、93 号和 94 号、162 号和 171 号、165 号和 166 号、168 号和 174 号以及 204 号和 205 号等 MEDV 矢量对间的模式距离为零, 说明它们的拓扑结构极为相似, MEDV 描述子不能分辨它们。

应用 MLR 方法建立 PAH 的色谱保留指数 RI 与 6 个 MEDV 描述子的 QSPR 方程^[30]：

$$\begin{aligned} RI = & (68.668\ 8 \pm 5.419\ 0) - (27.137\ 1 \pm 34.533\ 8)x_1 + \\ & (1.452\ 6 \pm 2.879\ 3)x_2 + (9.716\ 5 \pm 1.822\ 7)x_3 + \\ & (6.118\ 3 \pm 0.459\ 7)x_5 + (9.761\ 3 \pm 0.387\ 8)x_6 + \\ & (6.930\ 8 \pm 0.478\ 6)x_8 \end{aligned} \quad (4-26)$$

$$n = 209, m = 6, r = 0.981\ 3,$$

$$RMSE = 15.492, F = 876.964 \quad (\text{建模时})$$

由上式可知, 描述子 x_1 和 x_2 的回归系数标准偏差显然太大, 因而影响模型的稳定性。从相应的标准回归系数 (6 个描述子分别为 $-0.015823, 0.012137, 0.164351, 0.274203, 0.598964, 0.310604$) 看, 描述子 x_1 和 x_2 的标准回归系数也是最小的, 说明其对模型影响不是很大。因此, 可以预期从模型中删除这两个描

表 4-15 209 个芳烃与 6 个 MEDV-4 描述子及其实验气相色谱保留指数 RI

No	化合物	x_1	x_2	x_3	x_5	x_6	x_8	RI
1	1,2 - dihydronaphthalene	0	0	0	9.714 9	7.166 4	1.143 7	197.01
2	1,4 - dihydronaphthalene	0	0	0	9.894 8	6.962 2	1.143 7	197.01
3	tetralin	0	0	0	9.489 8	6.951 6	1.143 7	197.04
4	naphthalene	0	0	0	10.114 9	7.370 6	1.143 7	200.00
5	benzo[b] thiophene	0	0	0	7.688 5	6.530 0	1.143 7	201.47
6	indoline	0	0	0	8.423 0	7.119 0	1.143 7	204.74
7	indole	0	0	0	8.636 9	7.317 5	1.143 7	205.26
8	quinoline	0	0	0	10.621 5	7.730 7	1.143 7	209.70
9	isoquinoline	0	0	0	10.817 1	7.462 9	1.143 7	215.61
10	2 - methylnaphthalene	0	0.844 6	1.190 5	7.225 7	9.846 7	1.556 7	218.14
11	2 - methylbenzo[b] thiophene	0	0.683 8	1.209 8	5.474 1	8.279 5	1.608 6	218.74
12	azulene	0	0	0	10.196 2	7.370 6	1.143 7	219.95
13	quinoxaline	0	0	0	11.129 2	8.083 5	1.143 7	220.37
14	3 - methylbenzo[b] thiophene	0	0.675 2	1.388 5	5.707 4	7.081 5	2.573 4	221.02
15	1 - methylnaphthalene	0	0.761 5	1.388 5	7.946 6	8.109 2	2.573 4	221.04
16	8 - methylquinoline	0	0.789 0	1.388 5	8.386 0	8.536 6	2.573 4	223.02
17	1,2,3,4 - tetrahydroquinoline	0	0	0	10.042 2	7.360 8	1.143 7	225.97

续表

No	化合物	x_1	x_2	x_3	x_5	x_6	x_8	RI
18	6-甲基quinoline	0	0.854	11.190	5 7.717	0 10.222	2 1.556	7 229.82
19	1,2,2A,3,4,5, hexahydroacenaphthylene	0	0	0	8.057	1 11.048	3 4.107	5 232.70
20	biphenyl	0	0	0	13.542	0 7.767	7 1.143	7 233.96
21	2-ethylnaphthalene	0	1.425	40.360	1 8.070	3 11.037	2 1.556	7 236.08
22	1-ethylnaphthalene	0	1.409	90.432	8 8.708	1 9.497	7 2.573	4 236.56
23	3-ethylindole	0	0.713	91.388	5 6.547	7 7.977	0 2.573	4 236.66
24	2-methylindole	0	0.804	11.239	6 5.792	6 9.601	5 1.704	0 237.42
25	2,6- dimethylnaphthalene	0.022	41.627	22.443	0 4.382	3 12.231	4 2.015	5 237.58
26	2,7- dimethylnaphthalene	0.030	31.600	32.469	9 4.408	1 12.179	9 2.041	2 237.71
27	5-ethylbenzo[b] thiophene	0	1.400	80.360	1 5.662	7 10.143	0 1.556	7 238.46
28	2-methylbiphenyl	0	0.837	11.388	5 11.256	5 8.623	6 2.573	4 238.77
29	1,3- dimethylnaphthalene	0.066	81.363	42.821	7 5.343	4 10.013	5 3.272	3 240.25
30	1,4- naphthoquinone	0.086	62.040	04.771	4 5.905	4 8.593	7 4.130	1 240.82
31	1,7- dimethylnaphthalene	0.043	31.468	02.717	0 5.184	5 10.331	2 3.113	4 240.66
32	1,6- dimethylnaphthalene	0.030	31.517	22.667	9 5.128	9 10.442	4 3.057	8 240.72
33	2,2'- dimethylbiphenyl	0.043	31.536	12.915	0 9.098	1 9.225	3 4.130	1 241.94

续表

No	化合物	x_1	x_2	x_3	x_5	x_6	x_8	RI	
34	2,6 - dimethylquinoline	0.022	71.696	92.448	7	4.588	6 12.875	2 2.032	1 242.43
35	2,3 - dimethylnaphthalene	0.116	11.155	12.915	0	5.480	3 10.035	4 3.113	4 243.55
36	1,4 - dimethylnaphthalene	0.043	31.385	02.915	0	5.905	4 8.593	7 4.130	1 243.57
37	1,5 - dimethylnaphthalene	0.043	31.385	02.915	0	5.905	4 8.593	7 4.130	1 244.98
38	diphenylmethane	0	0	0	14.600	7 8.790	7 0.250	0 243.35	
39	acenaphthylene	0	0	0	8.717	0 11.961	3 4.288	9 244.63	
40	2,2' - bipyridyl	0	0	0	14.613	0 8.490	9 1.143	7 245.48	
41	1,2 - dimethylnaphthalene	0.116	11.072	13.113	0	6.201	1 8.297	9 4.130	1 246.49
42	1,8 - dimethylnaphthalene	0.066	81.280	33.019	7	6.064	2 8.276	0 4.288	9 249.52
43	2 - ethylbiphenyl	0	1.462	70.432	8	12.093	6 10.012	0 2.573	4 250.85
44	acenaphthene	0	0	0	8.483	1 11.552	9 4.288	9 251.29	
45	4 - methylbiphenyl	0	0.902	31.113	6	10.504	4 10.606	8 1.342	3 254.71
46	3 - methylbiphenyl	0	0.898	61.190	5	10.575	3 10.321	4 1.556	7 254.81
47	2,3 - dimethylindole	0.116	10.983	93.162	1	4.847	1 7.973	6 4.277	4 255.48
48	dibenzofuran	0	0	0	10.820	6 13.297	2 4.292	9 257.17	
49	2 - methyl - 1,4 - naphthoquinone	0.351	11.913	96.932	6	4.445	9 8.210	6 5.972	7 259.23

续表

No	化合物	x_1	x_2	x_3	x_5	x_6	x_8	RI
50	2,3,6 - trimethylnaphthalene	0.168 9 1.848	7 4.256 6	2.708 4	12.277 1	3.643 7	263.31	
51	1 - methylacenaphthylene	0	0.660 1 1.578	9 6.818 7	12.016 9	6.131 6	265.24	
52	2,3,5 - trimethylnaphthalene	0.189 7 1.689	5 4.530 6	3.510 6	10.377 0	4.741 6	265.90	
53	dibenzo - p - dioxin	0	0	0	12.470 5	16.455 3	3.123 8	267.27
54	fluorene	0	0	0	10.307 0	11.890 7	4.253 0	268.17
55	trans - 1,2,3,4,4a,9a - hexahydrodibenzothiphenene	0	0	0	9.518 4	10.560 4	3.839 8	269.67
56	cis - 1,2,3,4,4a,9a - hexahydrodibenzothiphenene	0	0	0	9.518 4	10.560 4	3.839 8	271.39
57	3,3' - dimethylbiphenyl	0.022 4 1.735	0 2.443 0	7.654 4	12.783 5	2.015 5	271.87	
58	9 - methylfluorene	0	0.585 8	1.732 2	9.149 2	10.514 4	6.787 1	272.38
59	2,3,5 - trimethylindole	0.189 7 1.578	4 4.579 7	2.189 4	10.019 7	4.888 9	272.57	
60	4,4' - dimethylbiphenyl	0.013 7 1.769	5 2.262 3	7.490 0	13.399 2	1.564 2	274.59	
61	5H - indeno[1,2 - b]pyridine	0	0	0	10.770 6	12.346 1	4.253 0	279.31
62	xanthene	0	0	0	11.887 9	15.048 7	3.083 9	280.48
63	9,10 - dihydroanthracene	0	0	0	11.312 2	13.610 5	3.019 6	284.89
64	9 - ethylfluorene	0	1.357 2	0.601 4	9.735 0	12.246 5	6.787 1	284.99
65	9,10 - dihydrophenanthrene	0	0	0	11.848 7	12.502 2	4.130 1	287.09

续表

No	化合物	x_1	x_2	x_3	x_5	x_6	x_8	R_I
66	1,2,3,4,5,6,7,8 - octahydroanthracene	0	0	0	10.253 2 13.453 2 3.113 4	287.69		
67	2 - methylfluorene	0	0.827 9 1.277 1 7.461 1 14.185 3 4.804 2	288.21				
68	1 - methylfluorene	0	0.721 0 1.522 1 8.255 9 12.268 9 5.925 8	289.03				
69	1,2,3,4,5,6,7,8 - octahydrophenanthrene	0	0	0	10.890 3 11.961 5 4.130 1	292.03		
70	1,2,3,4 - tetrahydridibenzothipene	0	0	0	9.556 1 10.710 1 4.182 0	294.30		
71	9 - fluorenone	0	0.875 4 2.869 2 9.160 2 10.611 1 7.148 2	294.79				
72	dibenzothipene	0	0	0	10.136 3 11.186 6 4.182 0	295.81		
73	1,2,3,4 - tetrahydrophenanthrene	0	0	0	11.492 4 12.436 1 4.130 1	297.21		
74	phenanthrene	0	0	0	12.096 5 12.910 6 4.130 1	300.00		
75	anthracene	0	0	0	11.488 2 14.352 3 3.113 4	301.69		
76	benzo[h]quinoline	0	0	0	12.577 6 13.366 8 4.130 1	302.22		
77	9,10 - dihydroacridine	0	0	0	11.631 4 14.402 3 3.059 0	304.33		
78	acridine	0	0	0	11.781 9 15.052 8 3.132 9	304.50		
79	1,2,3,4 - tetrahydrocarbazole	0	0	0	9.992 4 12.183 0 4.277 4	306.76		
80	phenanthridine	0	0	0	12.579 2 13.360 3 4.133 7	308.79		
81	benzo[f]quinoline	0	0	0	12.601 1 13.314 1 4.130 1	309.25		

续表

No	化合物	x ₁	x ₂	x ₃	x ₅	x ₆	x ₈	RI
82	carbazole	0	0	0	10.590 9 12.662 7 4.277 4	312.13		
83	9 - ethylcarbazole	0	1.371 70.683 7 9.774 1 12.916 6 7.561 9	313.97				
84	1 - phenyl-naphthalene	0	0	0	15.557 8 13.502 5 4.130 1	315.19		
85	1,2,3,10b - tetrahydrofluoranthene	0	0	0	11.094 5 14.703 5 8.727 0	316.37		
86	9 - n - propylfluorene	0	1.491 40.318 2 11.092 2 12.847 9 6.787 1	318.01				
87	3 - methylphenanthrene	0	0.860 51.304 0 9.211 2 15.184 4 4.741 6	319.46				
88	2 - methylphenanthrene	0	0.863 51.266 0 9.191 9 15.284 9 4.660 3	320.17				
89	3 - methylbenzo[f]quinoline	0	0.924 61.272 0 9.411 1 15.955 7 4.678 5	320.77				
90	2 - methylanthracene	0	0.863 51.266 0 8.583 7 16.726 6 3.643 7	321.57				
91	o - terphenyl	0	0	0	19.092 4 14.094 3 4.130 1	321.99		
92	4H - cyclopenta[def]phenanthrene	0	0	0	9.185 7 16.021 4 8.065 4	322.08		
93	9 - methylphenanthrene	0	0.760 21.578 9 10.025 4 13.139 0 5.972 7	323.06				
94	4 - methylphenanthrene	0	0.760 21.578 9 10.025 4 13.139 0 5.972 7	323.17				
95	1 - methylanthracene	0	0.777 51.502 0 9.323 8 14.888 6 4.741 6	323.33				
96	1 - methylphenanthrene	0	0.777 51.502 0 9.932 0 13.446 9 5.758 3	323.90				
97	2 - methylacridine	0	0.872 91.266 4 8.862 0 17.441 7 3.664 0	324.46				

续表

No	化合物	x_1	x_2	x_3	x_5	x_6	x_8	RI
98	9 - <i>n</i> - butylfluorene	0	1.535	50.199	312.583	513.166	16.787	1 328.99
99	9 - methylanthracene	0	0.677	11.776	910.138	012.843	25.972	7 329.13
100	4,5,9,10 - tetrahydropyrene	0	0	0	10.439	716.346	57.942	4 329.69
101	4,5 - dihydropyrene	0	0	0	10.678	316.771	17.942	4 330.01
102	thianthrene	0	0	0	10.918	512.134	92.829	2 330.13
103	anthrone	0	0.970	82.869	210.061	712.481	05.936	8 330.53
104	2 - phenylnaphthalene	0	0	0	14.881	515.068	92.899	0 332.59
105	9 - ethylphenanthrene	0	1.426	30.542	910.785	614.717	95.972	7 337.05
106	2 - ethylphenanthrene	0	1.443	30.412	910.055	416.550	94.660	3 337.50
107	3,6 - dimethylphenanthrene	0.022	41.659	02.670	16.371	617.366	65.399	0 337.83
108	2,7 - dimethylphenanthrene	0.013	71.691	82.567	16.310	717.612	55.213	9 339.23
109	1,2,3,6,7,8 - hexahydropyrene	0	0	0	10.074	916.295	77.942	4 339.38
110	6 - phenylquinoline	0	0	0	15.399	415.454	02.899	0 342.45
111	fluoranthene	0	0	0	11.525	415.325	19.117	9 344.01
112	9 - isopropylphenanthrene	0.250	00.852	53.085	810.025	413.899	27.551	6 345.78
113	1,8 - dimethylphenanthrene	0.022	41.492	83.066	17.813	313.891	67.432	2 346.26

续表

No	化合物	x_1	x_2	x_3	x_5	x_6	x_8	RI
1114	2-phenylindole	0	0	0	13.338 4 14.865 0	3.098 4	347.47	
1115	indeno[1,2,3-ij]isoquinoline	0	0	0	11.975 1 15.835 1	9.117 9	347.57	
1116	9-n-hexylfluorene	0	1.567 0 0.100 4	15.674 6 13.502 6	6.787 1	348.54		
1117	9-n-propylphenanthrene	0	1.528 8 0.289 7	12.211 9 15.260 9	5.972 7	350.30		
1118	pyrene	0	0	0	10.917 2 17.195 7	7.942 4	351.22	
1119	9,10-dimethylanthracene	0.043 3 1.216 1	3.691 9 8.914 9	11.080 0 8.959 1	355.49			
1120	benzo[lmn]phenanthridine	0	0	0	11.381 8 17.688 1	7.946 0	358.53	
1121	9-methyl-10-ethylphenanthrene	0.064 6 1.919 4	2.505 0 9.591 2	12.926 0 8.959 1	359.91			
1122	m-terphenyl	0	0	0	18.403 1 15.598 6	2.899 0	360.73	
1123	benzo[kl]xanthene	0	0	0	13.042 2 18.818 0	7.893 7	361.38	
1124	4H-benzo[def]carbazole	0	0	0	9.444 1 16.849 5	8.089 7	364.22	
1125	p-terphenyl	0	0	0	18.299 8 15.751 2	2.603 2	366.10	
1126	benzo[a]fluorene	0	0	0	12.363 0 17.267 3	7.620 8	366.74	
1127	11-methylbenzo[a]fluorene	0	0.586 7 1.915 0	11.289 6 15.418 1	10.543 3	367.04		
1128	9,10-dimethylphenanthrene	0.041 1 2.749 5	1.318 0 10.200 4	14.772 0 8.959 1	367.97			
1129	1-methyl-7-isopropylphenanthrene	0.277 0 1.606 5	4.402 4 7.059 2	16.598 2 7.609 2	368.67			

续表

No	化合物	x_1	x_2	x_3	x_5	x_6	x_8	RI
130	benzo[b]fluorene	0	0	0	11.727 7 18.786 3	6.559 5	369.39	
131	4 - methylpyrene	0	0.722 2 1.692 5	8.927 5	17.144 2 9.983 6	369.54		
132	2 - methylpyrene	0	0.824 6 1.418 4	8.145 8 19.105 2	8.804 4	370.15		
133	4,5,6 - trihydrobenz[de]anthracene	0	0	0	12.160 4 17.220 1	7.886 8	370.86	
134	1 - methylpyrene	0	0.739 4 1.615 6	8.834 0 17.452 1	9.769 2	373.55		
135	3,5 - diphenylpyridine	0	0	0	19.112 4 15.786 3	2.899 0	373.79	
136	5,12 - dihydronaphthacene	0	0	0	12.733 2 20.627 2	5.209 5	381.56	
137	9,10 - dimethyl - 3 - ethylphenanthrene	0.155 1 2.354 9	4.254 0 7.115 0	14.498 9 9.687 9	381.85			
138	9 - phenylcarbazole	0	0	0	16.608 0 16.854 0	9.666 3	382.09	
139	1 - ethylpyrene	0	1.407 9 0.580 2	9.573 5 19.067 7	9.769 2	385.35		
140	2,7 - dimethylpyrene	0.013 7 1.614 0	2.872 0 5.397 7	20.968 0 9.689 6	386.34			
141	1,2,3,4,5,6,7,8,9,10,11,12 - dodecahydrotriphenylene	0	0	0	13.160 5 15.029 8	8.959 1	386.36	
142	11 - benzo[a]fluorenone	0	0.867 7 3.145 1	11.299 7 15.508 4	10.944 7	386.41		
143	1,1' - binaphthyl	0	0	0	17.633 2 19.260 4	7.432 2	388.38	
144	benzo[b]naphtho[2,1 - d]thiophene	0	0	0	12.221 6 16.491 9	7.514 6	389.26	
145	benzo[ghi]fluoranthene	0	0	0	10.471 0 19.173 7	13.383 0	389.60	

续表

No	化合物	x_1	x_2	x_3	x_5	x_6	x_8	RI
146	benzo[c]phenanthrene	0	0	0	14.224 9 18.248 6	7.728 0	391.39	
147	benz[c]acridine	0	0	0	13.799 1 20.708 2	6.442 5	392.35	
148	9 – phenylanthracene	0	0	0	17.791 3 18.848 4	7.942 4	396.38	
149	cyclopenta[cd]pyrene	0	0	0	9.711 8 21.090 0	12.096 3	396.54	
150	benz[a]anthracene	0	0	0	13.529 4 19.915 5	6.415 6	398.50	
151	benz[a]acridine	0	0	0	13.821 4 20.658 8	6.437 4	398.74	
152	chrysene	0	0	0	14.137 6 18.473 8	7.432 2	400.00	
153	triphenylene	0	0	0	14.904 0 16.620 0	8.959 1	400.00	
154	benzo[a]carbazole	0	0	0	12.619 6 18.138 6	7.655 9	401.81	
155	1,2' – binaphthyl	0	0	0	16.934 5 20.872 0	6.080 1	405.35	
156	7 – benz[de]anthrene	0	0.946 0 3.066 8	11.297 5 15.845 8	11.000 3	406.54		
157	9 – phenylphenanthrene	0	0	0	17.723 2 18.973 1	7.728 0	406.90	
158	naphthacene	0	0	0	12.904 6 21.395 6	5.317 6	408.30	
159	benzo[b]carbazole	0	0	0	12.009 3 19.603 7	6.587 2	410.12	
160	11 – methylbenz[a]anthracene	0	0.796 3 1.577 6	11.349 5 20.349 9	8.161 0	412.72		
161	2 – methylbenz[a]anthracene	0	0.878 5 1.358 0	10.625 3 22.130 5	7.104 7	413.78		

续表

No	化合物	x_1	x_2	x_3	x_5	x_6	x_8	RI
162	1 - methylbenz[a]anthracene	0	0.779	0.1.654	5 11.442	9 20.042	1 8.375	5 414.37
163	1 - <i>n</i> - butylpyrene	0	1.549	20.209	1 12.496	4 19.969	1 9.769	2 414.87
164	1 - methyltriphenylene	0	0.776	1 1.692	5 12.836	7 16.646	1 11.000	3 416.32
165	9 - methylbenz[a]anthracene	0	0.879	7 1.306	5 10.606	8 22.252	8 7.000	9 416.50
166	3 - methylbenz[a]anthracene	0	0.879	7 1.306	5 10.606	8 22.252	8 7.000	9 416.63
167	9 - methyl - 10 - phenylphenanthrene	0	0.751	3 1.967	4 16.353	0 17.071	0 11.000	3 417.16
168	8 - methylbenz[a]anthracene	0	0.795	4 1.556	0 11.346	1 20.393	1 8.121	3 417.56
169	6 - methylbenz[a]anthracene	0	0.776	1 1.692	5 11.462	1 19.941	5 8.456	8 417.57
170	3 - methylchrysene	0	0.878	5 1.358	0 11.233	5 20.688	8 8.121	3 418.10
171	5 - methylbenz[a]anthracene	0	0.779	0 1.654	5 11.442	9 20.042	1 8.375	5 418.72
172	2 - methylchrysene	0	0.879	7 1.306	5 11.215	0 20.811	1 8.017	6 418.80
173	12 - methylbenz[a]anthracene	0	0.675	8 1.967	4 12.276	4 17.896	2 9.687	9 419.39
174	4 - methylbenz[a]anthracene	0	0.795	4 1.556	0 11.346	1 20.393	1 8.121	3 419.67
175	5 - methylchrysene	0	0.758	9 1.769	4 12.163	8 18.192	0 9.687	9 419.68
176	6 - methylchrysene	0	0.776	1 1.692	5 12.070	4 18.499	8 9.473	4 420.61
177	4 - methylchrysene	0	0.779	0 1.654	5 12.051	2 18.600	4 9.392	1 420.83

续表

No	化合物	x_1	x_2	x_3	x_5	x_6	x_8	RI
178	2,2' - biquinoline	0	0	0	16.900 9 23.903 4 4.832 9	421.12		
179	1 - phenylphenanthrene	0	0	0	17.633 4 19.162 0 7.432 2	421.66		
180	1 - methylchrysene	0	0.795 4 1.556 0 11.954 3 18.951 3 9.137 9	422.87				
181	7 - methylbenz[a]anthracene	0	0.693 0 1.890 5 12.183 0 18.204 1 9.473 4	423.14				
182	o - quaterphenyl	0	0	0	24.817 6 20.636 8 7.432 2	423.63		
183	2,2' - binaphthyl	0	0	0	16.249 3 22.443 8 4.786 9	423.91		
184	2, (2' - naphthyl) - benzo[b]thiophene	0	0	0	14.193 6 20.660 9 4.868 4	428.11		
185	1,3 - dimethyltriphenylene	0.066 8 1.412 7 3.314 8 10.221 9 18.246 2 12.015 0	432.32					
186	1,12 - dimethylbenz[a]anthracene	0.043 3 1.316 7 3.760 0 10.317 0 17.768 6 11.774 8	436.82					
187	benzo[j]fluoranthene	0	0	0	13.679 9 20.513 9 13.031 6	440.92		
188	benzo[b]fluoranthene	0	0	0	13.679 9 20.510 6 13.057 4	441.74		
189	benzo[k]fluoranthene	0	0	0	12.981 8 22.215 0 11.719 2	442.56		
190	7,12 - dimethylbenz[a]anthracene	0.043 3 1.230 7 3.996 0 11.057 1 15.930 6 12.872 8	443.38					
191	1,6,11 - trimethyltriphenylene	0.066 5 2.348 6 4.637 9 7.175 8 20.708 8 12.598 4	446.24					
192	dinaphtho[1,2 - b;1',2' - d]furan	0	0	0	14.975 5 24.212 6 11.470 8	450.20		
193	benzo[e]pyrene	0	0	0	13.763 1 20.734 9 13.168 6	450.73		

续表

No	化合物	x_1	x_2	x_3	x_5	x_6	x_8	RI
194	dibenzo[c, k]xanthene	0	0	0	15.084	824.518	511.474	8 451.57
195	benzo[a]pyrene	0	0	0	13.051	722.462	611.774	8 453.44
196	perylene	0	0	0	13.738	820.818	113.087	2 456.22
197	1,3,6,11 - tetramethyltriphenylene	0.173	02.877	46.368	1 4.638	5 22.153	813.690	7 461.72
198	3 - methylcholanthrene	0	0.724	21.642	610.620	722.324	013.321	2 468.44
199	<i>m</i> - quaterphenyl	0	0	0	23.379	023.618	4 4.786	9 472.81
200	indeno[1,2,3 - cd]pyrene	0	0	0	12.590	724.421	717.501	3 481.87
201	pentacene	0	0	0	14.346	428.511	3 7.632	0 486.81
202	<i>p</i> - quaterphenyl	0	0	0	23.143	623.872	9 4.135	9 488.18
203	dibenz[a, c]anthracene	0	0	0	16.396	423.648	111.560	4 495.01
204	dibenz[a, h]anthracene	0	0	0	15.598	825.552	5 9.850	4 495.45
205	benzo[b]chrysene	0	0	0	15.598	825.552	5 9.850	4 497.66
206	picene	0	0	0	16.207	124.110	810.867	0 500.00
207	benzo[ghi]perylene	0	0	0	12.681	124.607	617.693	9 501.32
208	dibenzo[def, mno]chrysene	0	0	0	12.024	826.209	416.433	2 503.89
209	2,3 - dihydrodibenzo[def, mno]chrysene	0	0	0	11.791	725.765	116.433	2 503.91

述子不会影响模型，且可能会增强模型的稳定性与统计显著性。
删除描述子 x_1 和 x_2 ，重建 4 参数模型：

$$RI = (69.670\ 9 \pm 5.223\ 6) + (9.522\ 7 \pm 1.053\ 9)x_3 +$$
$$(6.042\ 9 \pm 0.446\ 8)x_5 + (9.796\ 8 \pm 0.384\ 4)x_6 +$$
$$(6.915\ 7 \pm 0.474\ 7)x_8 \quad (4-27)$$
$$n=209, m=4, r=0.992\ 8, RMSE=15.533, F=1321.22$$

(建模时)

$$q=0.980\ 2, RMSV=15.982 \quad (\text{LOO 校验})$$

比较 4-27 式与 (4-26) 式可知, (4-27) 式虽然少了 2 个描述子, 但估计误差几乎没有变化, 相反统计显著性 (F) 增加了, 且模型中各个回归系数的标准偏差都有降低。从 LOO 分析也发现, 预测与估计能力相当。但是 $RMSE$ 与 $RMSV$ 都大于 15 的结果还是不太理想, 应当对模型做进一步改进。

若以 (4-27) 式进行 LOO 检验分析, 考察检验预测的保留指数 (RI_{LOO}) 与实验保留指数 RI_{OBS} 之间的误差大小(由 LOO 检验预测得到的保留指数 RI_{LOO} 与实验保留指数 RI_{OBS} 的相关情况如图 4-13A 所示), 可以发现有关的误差分布规律^[30]。首先, 多环芳烃中杂原子数增加, 误差增大。例如含 2 个氧原子的 53 号 PAH 和含 2 个硫原子的 102 号 PAH 的预测误差分别为 61.59 和 -56.68。含 1 个氧原子的 48 号样本及含 1 个硫原子的 70 号样本的预测误差分别为 38.21 和 -33.55。其次, 大多数含氧杂多环芳烃的预测误差为正, 而含硫杂原子的 PAH 则为负。在本文研究的 209 个 PAH 中, 12 个含氧 PAH 中就有 10 个误差为正, 其平均误差达 22.31。所有 11 个含硫 PAH 误差全为负, 其平均值达 -25.36。第三, 在 37 个含氮 PAH 中, 近一半 (18 个 PAH) 误差为负, 另一半为正, 平均误差仅 -3.50。这说明在这些杂原子混合的统一数据集中, 含氧化合物误差偏高, 含硫化合物误差偏低。这是 C-O 键减小与 O 的电负性增大同时增加了关于氢的 MEDV 描述子值,

而硫电性仅比碳稍大，但 C-S 键的键长比 C-C 键要长，较多地减少了 MEDV 描述子值所致。因此，我们采用拟碳方法使 C-O 键长加长到 C-C 键长，使 C-S 键长缩短到 C-C 键长，即 C-C，C-O 和 C-S 键的相对键长都视为 1。这样修正后得到的 MEDV 描述子再与保留指数 RI 相关，修正后的 4 参数方程为：

$$RI = (65.2361 \pm 4.8268) + (10.1766 \pm 0.9821)x_3 + (6.1822 \pm 0.4107)x_5 + (9.9837 \pm 0.3523)x_6 + (6.7888 \pm 0.4330)x_8 \quad (4-28)$$

$n = 209, m = 4, r = 0.9844, RMSE = 14.182, F = 1595.14$
(建模时)
 $q = 0.9834, RMSV = 14.625$ (LOO 预测)

考察 LOO 分析结果可知，有 37 个 PAH 的误差绝对值大于 20，其中有 17 个样本大于 25。这 17 个化合物分别是序号为 33, 48, 53, 70, 79, 82, 91, 102, 114, 119, 141, 143, 154, 159, 182, 184 和 200 号的 PAH 化合物。若删除这 17 个误差大的化合物，可得下列新模型：

$$RI = (56.6831 \pm 4.0217) + (11.6071 \pm 0.8176)x_3 + (7.0895 \pm 0.3564)x_5 + (9.7756 \pm 0.3564)x_6 + (6.8247 \pm 0.3581)x_8 \quad (4-29)$$

$n = 192, m = 4, r = 0.9907, RMSE = 11.146, F = 2464.38$
(建模时)
 $q = 0.9901, RMSV = 11.489$ (LOO 预测)

如果删除 37 个误差大于 20 的化合物，则得模型如下：

$$RI = (58.0427 \pm 3.5878) + (11.1161 \pm 0.7712)x_3 + (7.0236 \pm 0.3180)x_5 + (9.7877 \pm 0.2542)x_6 + (6.9799 \pm 0.3191)x_8 \quad (4-30)$$

$n = 172, m = 4, r = 0.9936, RMSE = 9.212, F = 3222.40$

(建模时)

$$q = 0.9936, \text{RMSV} = 9.518$$

(LOO 预测)

由 (4-30) 式估计的保留指数 (RI_{172}) 对实验保留指数 (RI_{OBS}) 作图如图 4-13B。

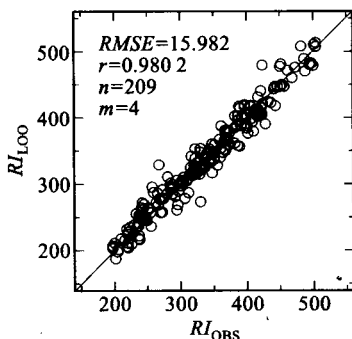


图 4-13A 209 个 PAH 预测与实验保留指数相关图

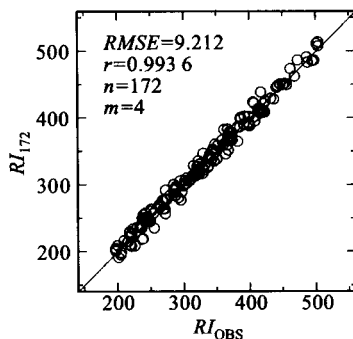


图 4-13B 172 个 PAH 估计与实验保留指数相关图

参考文献

- [1] Balban A T. High discrimination distance - based topological index. Chem. Phys. Lett., 1982, 89: 399~404
- [2] Hosoya H. Topological index. A newly proposed quantity characterising the topological nature of structural isomers of saturated hydrocarbons. Bull. Chem. Soc. Japan, 1971, 44: 2332~2339
- [3] Randic M. On characterization of molecular Branching. J. Am. Chem. Soc., 1975, 97: 6609~6615
- [4] Wiener H. Structural determination of paraffin boiling points. J. Am. Chem. Soc., 1947, 69: 17~20
- [5] Liu S S, Cai S X, Cao C Z, Li Z L. Molecular electronegative distance vector (MEDV) relating to 15 properties of alkanes. I. Chem. Inf. Comput. Sci., 2000, 40(6): 1337~1348

- [6] 许禄, 化学计量学方法北京: 科学出版社, 1995
- [7] Basak S C, Grunwald G D. Molecular similarity and estimation of molecular properties. *J. Chem. Inf. Comput. Sci.*, 1995, 35: 366~372
- [8] Egolf L M, Jurs P C. Prediction of boiling points of organic heterocyclic compounds using regression and neural network techniques. *J. Chem. Inf. Comput. Sci.*, 1993, 33: 616~625
- [9] Egolf L M, Wessel M D, Jurs P C. Prediction of boiling points and critical temperatures of industrially important organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.*, 1994, 34: 947~965
- [10] Stanton D T, Egolf L M, Jurs P C, Hicks M G. Computer - assisted prediction of normal boiling points of pyrans and pyrroles. *J. Chem. Inf. Comput. Sci.*, 1992, 32: 306~316
- [11] Wessel M D, Jurs P C. Prediction of normal boiling points for a diverse set of industrially important organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.*, 1995, 35: 841~850
- [12] Wessel M D, Jurs P C. Prediction of normal boiling points of hydrocarbons from molecular structure. *J. Chem. Inf. Comput. Sci.*, 1995, 35: 68~76
- [13] Mihalic Z, Nikolic S, Trinajstic N. Comparative study of molecular descriptors derived from the distance matrix. *J Chem Inf. Comput Sci.*, 1992, 32: 28~37
- [14] Katritzky A R, Maran U, Lobanov V S, Karelson M. Perspective: structurally diverse quantitative structure - property relationship correlations of technologically relevant physical properties. *J. Chem. Inf. Comput. Sci.*, 2000, 40(1): 1~18
- [15] Liu S S, Cao C Z, Li Z L. Approach to estimation and prediction for normal boiling point (NBP) of alkanes based on a novel molecular distance - edge (MDE) vector, λ . *J. Chem. Inf. Comput. Sci.*, 1998, 38(3): 387~394
- [16] Gakh A A, Gakh E G, Sumpter B G, Noid D W. Neural network - graph theory approach to the prediction of the physical properties of organic compounds. *J. Chem. Inf. Comput. Sci.*, 1994, 34(4): 832~839
- [17] Needham D E, Wei I - C, Seybold P G. Molecular modeling of the physi-

- cal properties of the alkanes. J. Am. Chem. Soc., 1988, 110(13): 4186~4194
- [18] Somayajulu G R, Zwolinski B J. Generalized treatment of alkanes. 2. J. Chem. Soc. – Faraday Trans. II, 1972, 68(11): 1971~1987
- [19] Kier L B, Hall L H. Molecular connectivity analysis of structure influencing chromatographic retention indexes. J. Pharm. Sci., 1979, 68: 120~122
- [20] Michotte Y, Massart D L. Molecular connectivity and retention indexes. J. Pharm. Sci., 1977, 66: 1630~1632
- [21] Millership J S, Woolfson A D. The retention between molecular connectivity and gas chromatographic retention data. J. Pharm. Pharmacol., 1978, 30: 483~485
- [22] Randic M. The structure origin of chromatographic retention data. J. Chromatogr., 1978, 161: 1~14
- [23] 武杰, 陆婉珍. 开管柱程序升温色谱对直馏汽油中烃的分析. 分析化学, 1984, 12(7): 572~578
- [24] Liu S S, Liu H L, Xia Z N, Cao C Z, Li Z L. Molecular distance – edge vector (μ) and chromatographic retention index of alkanes. J. Chin. Chem. Soc., 2000, 47(3): 455~460
- [25] 尹情胜, 刘树深, 李志良. 烯烃的分子距离边数矢量及沸点的估计与预测. 广西科学, 1998, 5(4): 291~293
- [26] 徐寿昌. 有机化学. 北京: 高等教育出版社, 1981
- [27] 李志良, 曹晨忠. 有效碳链长度与脂肪醇的沸点. 有机化学, 1998, 18(4): 360~364
- [28] Liu S S, Liu H L, Xia Z N, Cao C Z, Li Z L. Molecular distance – edge vector (μ): an extension from alkanes to alcohols. J. Chem. Inf. Comput. Sci., 1999, 39(6): 951~957
- [29] Lee M L, Vassilaros D L, White C M, Novotny M. Retention indices for programmed – temperature capillary – column gas chromatography of polycyclic aromatic hydrocarbons. Anal. Chem., 1979, 51(6): 768~773
- [30] Liu S S, Yin C S, Cai S X, Li Z L. Molecular structural vector description and retention index of polycyclic aromatic hydrocarbons. Chemom. Intel. Lab. Syst., 2002, 61: 3~15

第5章

生物活性物质 QSAR 研究

包括药物分子在内的生物活性物质的分子结构表征是现代分子设计与新分子创制特别是药物分子设计与新药创制中不可缺少的重要手段之一。众所周知,大量药物分子是由碳、氢元素组成且含有氧、氮、硫和卤素等多种杂原子的有机分子体系。对这些原子及其构成的分子化学环境进行定量表征是非常有实用价值的。

前一章通过多元线性回归 (MLR) 方法对基于 4 种原子类型的分子电性距离矢量 MEDV-4 与烷烃的多种物理化学性质及热力学函数如沸点、密度、折光率、热焓、保留指数等的定量结构-性质关系 (QSPR) 以及芳烃、醇、多环芳烃等沸点与保留指数的 QSPR 研究作了比较深入的讨论。在这些化合物中,主要是含碳及少数几个杂原子的较为简单的有机分子体系而不是真正的药物分子体系。对于高度原子多样性和多种化学键类型的生物活性分子比如药物分子体系, MEDV-4 分子描述子对其结构的分辨率会降低,因而必须扩展 MEDV-4。为此,作者又提出了基于 13 种原子类型,不使用相对电性只使用各种化学键相对距离的分子全息距离矢量 (molecular holographic distance vector 简称 MHDV)^[1, 2],以及基于 13 种原子类型,以改进的电拓扑状态指数代替相对电性、

拓扑距离代替相对距离的分子电性距离矢量 MEDV-13^[3~5]。这两个新的矢量分子描述子可以对药物分子中有关原子多样性进行比较充分的表征。本章将应用主成分回归 (principal component regression 简称 PCR) 方法和偏最小二乘回归 (partial least square regression 简称 PLSR) 首先讨论 MHDV 矢量分子描述子应用于两组二肽和一组生物可降解性取代苯化合物的定量结构-生物活性相关 quantitative structure-activity relationship, QSAR 研究, 然后介绍 MEDV-13 矢量分子描述子在两组药物分子生物活性的 QSAR 研究中的应用, 从而说明基于 13 种原子类型的分子电性距离矢量 MHDV 与 MEDV-13 应用于生物活性物质分子结构表征的有效性。

5.1 MHDV 用于两组二肽分子 QSAR 研究

肽分子是由 20 个基本天然氨基酸按一定次序连接而成的高度柔性的和原子多样性的分子。一个肽分子即使是最简单的二肽分子所含杂原子的种类和个数也比前面的 209 个多环芳烃分子要复杂, 用 MEDV-4 分子描述子来表征其结构较为困难。因此, 本章采用分子全息距离矢量 MHDV 来表征二肽分子的结构。由于 MHDV 矢量中描述子较多 (最多可达 91 个) 同时为了便于分析生物活性的影响因素, 本章将采用第三章提出的主成分回归即 PCR 方法构建定量药物分子结构与生物活性的相关模型。同时为了获得只含少数几个主要影响因素的习用 MLR 方程, 需要对描述子变量进行合理选择, 删除显著性差的描述子之后, 还需要根据各个分子体系中的 MEDV 描述子变量矩阵的主成分分析结果, 对载荷矢量进行分析, 以确定影响主成分的原始描述子, 进而经矩阵变换给出这些分子描述子对应的习用回归方程或似回归方程 (MLR-like model)。

5.1.1 二肽分子 MHDV 描述子

最近有作者完成了两组二肽分子的 QSAR 研究。这两组二肽分子是 Zaliani 和 Collantes 等作者^[6, 7]研究使用的 48 个苦味活性肽(bitter tasting thresholds 简称 BT, 其活性用其浓度的负对数即 pT 表示)以及由 Zaliani 和 Cocchi 等作者^[7, 8]研究使用的 58 个抑制血管紧张素转移酶肽(angiotensin-converting enzyme, 简称 ACE 其活性以 pIC₅₀ 表示)。他们首先利用 20 种天然氨基酸的各种性质包括物理化学性质进行主成分分析;进而推导一组复杂的 3D 描述子, 比如 z-scores^[6, 9], t-scores^[8] 和 MS-WHIM^[7, 10] 描述子;最后再通过多元线性回归等多变量分析技术建立相应的 QSAR 方程。作者直接选择 Zaliani 报告^[7]中给出的两组二肽化合物及其生物活性数据作为研究对象。这些二肽分子的结构以单字母缩写的氨基酸符号按顺序书写, 并按俗成的方法将 N-末端写在左边而 C-末端写在右边。例如图 5-1 所示的二肽分子可缩写成 LW。

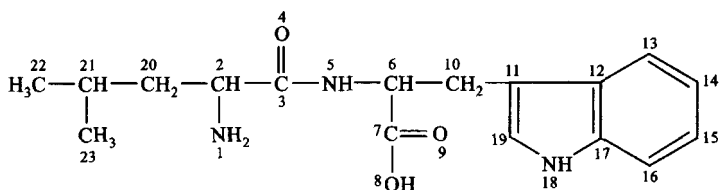


图 5-1 二肽分子(H₂N-Leucine-Tryptophan-COOH)的分子结构

分子全息距离矢量 MHDV 是基于 13 个原子类型提出的。全息距离就是包括这些原子类型所对应的各个原子属性的全部距离。原子类型与原子属性定义见第二章。在二肽分子或其他多肽分子中原子属性有 10 个(参见表 2-3), 即“C1”, “CC”, “C2”, “N1”, “NN”, “N2”, “O1”, “O2”, “S1”, “NP”等。其中 NP 表示二肽分子特殊肽键中的氮原子, 那么 C2-NP 就表示肽键。

按照第二章的定义, MHDV 不考虑相对电性, 因而其描述子可写成:

$$x_r = h_{kl} = \sum_{i \in k, j \in l} \frac{1}{d_{ij}^2} \quad (k = 1, 2, 3, \dots, 13; k \leq l \leq 13) \quad (2-4)$$

式中 k 或 l 是各非氢原子的原子类型; 原子 i 和 j 分别属于第 k 原子类型和第 l 原子类型; d_{ij} 表示原子 i 和原子 j 之间的相对距离, 是从原子 i 到原子 j 的各个路径中化学键相对键长加和之最小值。加和号“ $\sum$ ”表示对所有属于原子类型 k 的非氢原子 i 和属于原子类型 l 的非氢原子 j 求和。

下面以上述示例二肽 LW 分子为例说明 MHDV 描述子的计算过程:

LW 中含有 23 个非氢原子, 其原子属性和相应原子类型分别如表 5-1A 所示。由表 5-1A 可知, 二肽 LW 中缺少第 4, 7, 8, 10, 11, 12 和 13 原子类型, 因而与这 7 个原子类型相关的 MHDV 描述子共 70 个(根据表 2-7 的原子类型序号与 MHDV 描述子序号之间的相互关系而得)的值都等于 0。另外 LW 分子中只有一个非氢原子是第 5 原子类型, 导致 $x_{47} = h_{5,5} = 0$ 因而只余下 20 个 MHDV 描述子的取值不为 0。这些描述子是: $x_1, x_2, x_3, x_5, x_6, x_9, x_{14}, x_{15}, x_{17}, x_{18}, x_{21}, x_{26}, x_{28}, x_{29}, x_{32}, x_{48}, x_{51}, x_{56}, x_{59}, x_{77}$ 。要计算这些描述子的值, 需要由各个原子属性表达的化学键的键长数据, 对于二肽分子, 相对键长的数据如表 5-1B 所示。有了相对键长数据, 就可求得两两原子之间的最短相对距离, 结果如表 5-1C 所示。

表 5-1A 二肽分子 LW 中 23 个非氢原子的原子属性与原子类型

原子编号	原子属性	原子类型	原子编号	原子属性	原子类型
1	"N1"	5	2	"C1"	3

续表

原子编号	原子属性	原子类型	原子编号	原子属性	原子类型
3	"C2"	3	14	"CC"	2
4	"O2"	9	15	"CC"	2
5	"NP"	6	16	"CC"	2
6	"C1"	3	17	"CC"	3
7	"C2"	3	18	"N1"	6
8	"O1"	9	19	"CC"	2
9	"O2"	9	20	"C1"	2
10	"C1"	2	21	"C1"	3
11	"CC"	3	22	"C1"	1
12	"CC"	3	23	"C1"	1
13	"CC"	2			

表 5-1B 二肽分子中各种化学键的相对键长数据

二肽中的化学键	键长/nm	相对键长
C1 - C1, C1 - CC, C1 - C2	0.154	1.0000
C1 - O1, CC - O1, C2 - O1	0.143	0.9286
C1 - N1, CC - N1, C2 - N1, C1 - NP	0.147	0.9545
CC - CC, CC - C2	0.144	0.9351
C1 - S1, CC - S1	0.182	1.1818
CC - NN	0.141	0.9156
C2 - NP	0.133	0.8636
C2 - N2	0.130	0.8442
C2 - O2	0.122	0.7922

表 5-1C LW 中 23 个非氢原子的最短相对距离

i	j	$d(i,j)$	i	j	$d(i,j)$	i	j	$d(i,j)$	i	j	$d(i,j)$
1	2	0.9545	2	4	1.7922	3	7	2.8182	4	11	4.6104
1	3	1.9545	2	5	1.8636	3	8	3.6104	4	12	5.5455
1	4	2.7468	2	6	2.8182	3	9	3.7468	4	13	6.4805
1	5	2.8182	2	7	3.8182	3	10	2.8182	4	14	7.4156
1	6	3.7727	2	8	4.6104	3	11	3.8182	4	15	8.3506
1	7	4.7727	2	9	4.7468	3	12	4.7532	4	16	7.4156
1	8	5.5649	2	10	3.8182	3	13	5.6883	4	17	6.4805
1	9	5.7013	2	11	4.8182	3	14	6.6234	4	18	6.5000
1	10	4.7727	2	12	5.7532	3	15	7.5584	4	19	5.5455
1	11	5.7727	2	13	6.6883	3	16	6.6234	4	20	2.7922
1	12	6.7078	2	14	7.6234	3	17	5.6883	4	21	3.7922
1	13	7.6429	2	15	8.5584	3	18	5.7078	4	22	4.7922
1	14	8.5779	2	16	7.6234	3	19	4.7532	4	23	4.7922
1	15	9.5130	2	17	6.6883	3	20	2.0000	5	6	0.9545
1	16	8.5779	2	18	6.7078	3	21	3.0000	5	7	1.9545
1	17	7.6429	2	19	5.7532	3	22	4.0000	5	8	2.7468
1	18	7.6623	2	20	1.0000	3	23	4.0000	5	9	2.8831
1	19	6.7078	2	21	2.0000	4	5	1.6558	5	10	1.9545
1	20	1.9545	2	22	3.0000	4	6	2.6104	5	11	2.9545
1	21	2.9545	2	23	3.0000	4	7	3.6104	5	12	3.8896
1	22	3.9545	3	4	0.7922	4	8	4.4026	5	13	4.8247
1	23	3.9545	3	5	0.8636	4	9	4.5390	5	14	5.7597
2	3	1.0000	3	6	1.8182	4	10	3.6104	5	15	6.6948

续表

i	j	$d(i, j)$	i	j	$d(i, j)$	i	j	$d(i, j)$	i	j	$d(i, j)$
5	16	5.7597	6	22	5.8182	8	14	6.5974	9	23	7.7468
5	17	4.8247	6	23	5.8182	8	15	7.5325	10	11	1.0000
5	18	4.8442	7	8	0.7922	8	16	6.5974	10	12	1.9351
5	19	3.8896	7	9	0.9286	8	17	5.6623	10	13	2.8701
5	20	2.8636	7	10	2.0000	8	18	5.6818	10	14	3.8052
5	21	3.8636	7	11	3.0000	8	19	4.7273	10	15	4.7403
5	22	4.8636	7	12	3.9351	8	20	5.6104	10	16	3.8052
5	23	4.8636	7	13	4.8701	8	21	6.6104	10	17	2.8701
6	7	1.0000	7	14	5.8052	8	22	7.6104	10	18	2.8896
6	8	1.7922	7	15	6.7403	8	23	7.6104	10	19	1.9351
6	9	1.9286	7	16	5.8052	9	10	2.9286	10	20	4.8182
6	10	1.0000	7	17	4.8701	9	11	3.9286	10	21	5.8182
6	11	2.0000	7	18	4.8896	9	12	4.8636	10	22	6.8182
6	12	2.9351	7	19	3.9351	9	13	5.7987	10	23	6.8182
6	13	3.8701	7	20	4.8182	9	14	6.7338	11	12	0.9351
6	14	4.8052	7	21	5.8182	9	15	7.6688	11	13	1.8701
6	15	5.7403	7	22	6.8182	9	16	6.7338	11	14	2.8052
6	16	4.8052	7	23	6.8182	9	17	5.7987	11	15	3.7403
6	17	3.8701	8	9	1.7208	9	18	5.8182	11	16	2.8052
6	18	3.8896	8	10	2.7922	9	19	4.8636	11	17	1.8701
6	19	2.9351	8	11	3.7922	9	20	5.7468	11	18	1.8896
6	20	3.8182	8	12	4.7273	9	21	6.7468	11	19	0.9351
6	21	4.8182	8	13	5.6623	9	22	7.7468	11	20	5.8182

续表

<i>i</i>	<i>j</i>	<i>d(i,j)</i>	<i>i</i>	<i>j</i>	<i>d(i,j)</i>	<i>i</i>	<i>j</i>	<i>d(i,j)</i>	<i>i</i>	<i>j</i>	<i>d(i,j)</i>
11	21	6.8182	13	18	2.8247	15	18	2.8247	17	22	9.6883
11	22	7.8182	13	19	2.8052	15	19	3.7792	17	23	9.6883
11	23	7.8182	13	20	7.6883	15	20	9.5584	18	19	0.9545
12	13	0.9351	13	21	8.6883	15	21	10.5584	18	20	7.7078
12	14	1.8701	13	22	9.6883	15	22	11.5584	18	21	8.7078
12	15	2.8052	13	23	9.6883	15	23	11.5584	18	22	9.7078
12	16	1.8701	14	15	0.9351	16	17	0.9351	18	23	9.7078
12	17	0.9351	14	16	1.8701	16	18	1.8896	19	20	6.7532
12	18	1.8896	14	17	2.8052	16	19	2.8442	19	21	7.7532
12	19	1.8701	14	18	3.7597	16	20	8.6234	19	22	8.7532
12	20	6.7532	14	19	3.7403	16	21	9.6234	19	23	8.7532
12	21	7.7532	14	20	8.6234	16	22	10.6234	20	21	1.0000
12	22	8.7532	14	21	9.6234	16	23	10.6234	20	22	2.0000
12	23	8.7532	14	22	10.6234	17	18	0.9545	20	23	2.0000
13	14	0.9351	14	23	10.6234	17	19	1.9091	21	22	1.0000
13	15	1.8701	15	16	0.9351	17	20	7.6883	21	23	1.0000
13	16	2.8052	15	17	1.8701	17	21	8.6883	22	23	2.0000
13	17	1.8701									

利用表 5-1A 的原子类型和表 5-1C 的最短相对距离就可依据(2-4)式计算各个非零 MHDV 描述子:

$$x_1 = h_{1,1} = \frac{1}{2^2} = 0.2500$$

$$x_2 = h_{1,2} = \frac{2}{2^2} + \frac{2}{6.8182^2} + \frac{2}{8.7532^2} + \frac{2}{9.6883^2} + \frac{4}{10.6234^2} +$$

$$\frac{2}{11.5584^2} = 0.6408$$

$$x_3 = h_{1,3} = \frac{2}{1^2} + \frac{2}{3^2} + \frac{2}{4^2} + \frac{2}{5.8182^2} + \frac{2}{6.8182^2} + \frac{2}{7.8182^2} + \frac{2}{8.7532^2} + \frac{2}{9.6883^2} = 2.5295$$

$$x_5 = h_{1,5} = \frac{2}{3.9545^2} = 0.1279$$

$$x_6 = h_{1,6} = \frac{2}{4.8636^2} + \frac{2}{9.7078^2} = 0.1058$$

$$x_9 = h_{1,9} = \frac{2}{4.7922^2} + \frac{2}{7.7468^2} + \frac{2}{7.6104^2} = 0.1549$$

同理，根据表 5-1A 与 5-1C 给出的数据可计算出其他 14 个 MHDV 描述子值，其结果总结表 5-1D 中。

表 5-1D 二肽分子 LW 的 20 个非零 MHDV 描述子的计算结果

i	x_i	i	x_i	i	x_i	i	x_i	i	x_i
1	0.2500	6	0.1058	17	0.3832	28	1.6573	51	0.1956
2	0.6408	9	0.1549	18	2.4109	29	5.1102	56	0.0426
3	2.5295	14	5.2131	21	0.8916	32	6.2010	59	0.7018
5	0.1279	15	12.0807	26	6.5715	48	0.1429	77	0.4378

由上述示例可知，要计算 MHDV 描述子，最短距离计算是关键，对于非氢原子较多的分子手工完成是相当困难的。本文应用第二章提出的 GITMHDV 程序计算全部 48 个二肽及 58 个二肽的 91 个 MHDV 矢量描述子。

计算结果表明在 48 个二肽分子体系中，实际上只有 26 个非零 MHDV 描述子（编号名称及描述子数值参见表 5-2A），其余 65 个 MHDV 描述子对所有 48 个二肽均为零。在这 26 个非零 MHDV 描述子中有 6 个 MHDV 描述子仅有 1 (x_7) 或 3 个样本 ($x_{19}, x_{30}, x_{47}, x_{49}, x_{66}$) 取值不为零，这么少的非零样本数均小于

总样本的 10%) 在统计上显然是不显著的, 进行主成分分析并建立 QSAR 模型时, 应将其删除, 以提高模型的统计显著性。因此, 最终进入 PCR 分析步骤的 MHDV 描述子实际上只有 20 个。

对于 58 个二肽, 只有 33 个 MHDV 描述子的取值不全为零 (编号参见表 5-2B) 其余 58 个 MHDV 描述子对所有 58 个二肽均为零。在这 33 个非零 MHDV 描述子中, 有 8 个描述子只有 1 个 (x_{57}) 或 2 个 ($x_{10}, x_{22}, x_{33}, x_{52}, x_{60}, x_{78}$) 或 3 个 (x_7) 样本非零值, 从统计显著性考虑应予删除。因此, 进入 58 肽组 PCR 分析的 MHDV 描述子实际上只有 25 个。

为了说明 MHDV 描述子对这些二肽分子结构的分辨能力, 需要计算各个样本 MHDV 矢量描述子之间的模式距离, 该模式空间的维数由非零 MHDV 描述子个数决定。研究结果表明, 对于 48 肽组在 1 128 个模式距离中, 其最小距离为 0.2562 而最大距离为 24.0256 说明 MHDV 能完全分辨 48 个肽的分子结构 其距离分布见图 5-2A。对于 58 肽组在 1 653 个距离中, 最小值为 0.1276 最大值为 16.377, 也说明 MHDV 能完全分辨 58 个肽的分子结构, 其距离分布如图 5-2B。两组二肽的 MHDV 矢量模式空间距离范围与频率分布如表 5-3 所示。

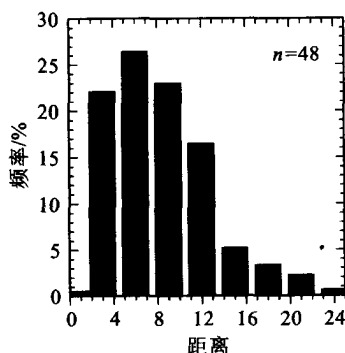


图 5-2A 48 肽组 MHDV 空间距离分布图

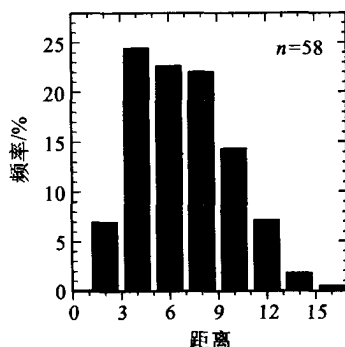


图 5-2B 58 肽组 MHDV 空间距离分布图

表 5-2A 48 肽组 20 个 MHDV 描述子

No	二肽	x_1	x_2	x_3	x_5	x_6	x_9	x_{14}	x_{15}	x_{17}	x_{18}
1	GV	0.2500	0.0862	2.8594	0.0600	0.2291	0.3628	0.0000	1.2631	1.0975	0.2879
2	GL	0.2500	0.5591	2.4334	0.0436	0.1279	0.2330	0.0686	3.6135	1.1414	0.5497
3	GI	0.1111	1.3226	1.8964	0.0518	0.1785	0.2979	0.0431	2.6928	1.1275	0.4025
4	GP	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	2.5131	3.7525	1.2556	0.0000
5	GF	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	7.0096	7.4251	1.2337	0.7979
6	GW	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	5.2626	10.6816	1.2190	2.5823
7	GY	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	4.0641	10.1478	1.2201	0.7678
8	AV	0.3091	0.0000	4.3503	0.3218	0.3511	0.5531	0.0000	0.0000	0.0000	0.0000
9	AL	0.2930	0.5431	3.8837	0.3054	0.2498	0.4233	0.0000	2.4445	0.0439	0.2618
10	AF	0.0000	0.1342	1.3912	0.2618	0.1219	0.1903	6.8187	6.3785	0.1362	0.5100
11	VG	0.2500	0.0862	2.7813	0.2291	0.1340	0.2288	0.0000	1.4970	0.0703	1.0975
12	VA	0.3091	0.0000	4.3550	0.2730	0.3957	0.5504	0.0000	0.0000	0.0000	0.0000
13	VV	0.5860	0.0000	5.9312	0.2891	0.3631	0.5915	0.0000	0.0000	0.0000	0.0000
14	VL	0.5654	0.5591	5.4460	0.2727	0.2619	0.4617	0.0000	2.4876	0.0439	0.2618

续表

No	二肽	x_1	x_2	x_3	x_5	x_6	x_9	x_{14}	x_{15}	x_{17}	x_{18}
15	LG	0.2500	0.5591	2.3902	0.1279	0.0845	0.1549	0.0686	3.7646	0.3320	1.2195
16	LA	0.2930	0.5431	3.9234	0.1718	0.3463	0.4765	0.0000	2.3617	0.2618	0.1219
17	LL	0.5514	1.0860	5.0072	0.1715	0.2124	0.3879	0.0431	4.8653	0.3057	0.3837
18	LF	0.2500	0.6556	2.4820	0.1279	0.0845	0.1549	6.9529	8.8698	0.3979	0.6320
19	LW	0.2500	0.6408	2.5295	0.1279	0.1058	0.1549	5.2131	12.0807	0.3832	2.4109
20	LY	0.2500	0.6378	2.4998	0.1279	0.0845	0.1699	4.0111	11.5607	0.3844	0.6018
21	IG	0.1111	1.3226	1.8358	0.1785	0.1093	0.1919	0.0431	2.8877	0.1848	1.1645
22	IA	0.1622	1.2795	3.3960	0.2224	0.3710	0.5134	0.0000	1.4337	0.1146	0.0670
23	IV	0.4369	1.2930	4.9641	0.2385	0.3384	0.5546	0.0000	1.4633	0.1146	0.0670
24	IL	0.4196	1.8338	4.4817	0.2221	0.2372	0.4248	0.0295	3.9428	0.1585	0.3288
25	II	0.2893	2.5757	3.9795	0.2303	0.2878	0.4897	0.0215	2.9656	0.1446	0.1815
26	IP	0.1111	1.4247	1.9084	0.1785	0.0000	0.1919	2.3522	4.4095	0.2726	0.0000
27	IW	0.1111	1.4104	1.9999	0.1785	0.1331	0.1919	5.1833	11.1655	0.2360	2.3523
28	IN	0.1111	1.3011	1.9463	0.2080	0.1093	0.2226	0.0295	3.9428	0.4368	0.3288

续表

No	二肽	x_1	x_2	x_3	x_5	x_6	x_9	x_{14}	x_{15}	x_{17}	x_{18}
29	ID	0.1111	1.3011	1.9463	0.1785	0.1093	0.2524	0.0295	3.9428	0.1585	0.3288
30	IQ	0.1111	1.3389	1.9376	0.2020	0.1093	0.2162	1.0511	4.6900	0.5778	0.4433
31	IE	0.1111	1.3389	1.9376	0.1785	0.1093	0.2398	1.0511	4.6900	0.1885	0.4433
32	IK	0.1111	1.3914	1.9084	0.1976	0.1093	0.1919	3.6914	3.8613	1.7751	0.5480
33	IS	0.1111	1.3011	1.9084	0.1785	0.1093	0.2305	0.0295	2.9213	0.1585	0.3288
34	IT	0.1490	1.2715	3.4618	0.2085	0.2238	0.6807	0.0000	1.4633	0.1146	0.0670
35	PA	0.0000	0.1031	1.4445	0.0000	0.3057	0.3216	2.2500	2.2632	0.0000	1.8802
36	PL	0.2500	0.6092	2.4925	0.0000	0.1715	0.2330	2.3531	4.7808	0.0000	2.1859
37	PI	0.1111	1.3778	1.9690	0.0000	0.2303	0.2979	2.3231	3.8391	0.0000	2.0248
38	PY	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	6.4420	11.4383	0.0000	2.4827
39	PF	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	9.4056	8.7148	0.0000	2.5264
40	FG	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	7.0096	7.4942	0.5803	1.3896
41	FL	0.2500	0.6556	2.5252	0.0436	0.1279	0.2330	6.9529	8.7878	0.5539	0.5539
42	FP	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	9.5280	9.1369	0.6681	0.0000

续表

No	二肽	x_1	x_2	x_3	x_5	x_6	x_9	x_{14}	x_{15}	x_{17}	x_{18}
43	FF	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	14.1106	12.7924	0.6462	0.8021
44	FY	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	11.1293	15.5228	0.6326	0.7720
45	WE	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	6.3031	12.7737	0.5273	2.4831
46	WW	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	10.5744	19.4164	0.5749	4.4651
47	YL	0.2500	0.6378	2.5429	0.0436	0.1279	0.2480	4.0111	11.4881	0.5238	0.5314
48	SL	0.2500	0.5430	2.4925	0.0436	0.1279	0.2663	0.0431	3.8357	0.3057	0.3837

表 5-2A 48 肽组 20 个 MHDV 描述子(续)

No	x_{21}	x_{26}	x_{28}	x_{29}	x_{32}	x_{48}	x_{51}	x_{56}	x_{59}	x_{77}
1	0.4028	2.8043	0.4198	2.9618	5.6197	0.1259	0.1956	0.0000	0.6176	0.4378
2	0.7243	1.8581	0.4059	2.8145	5.4795	0.1259	0.1956	0.0000	0.6176	0.4378
3	0.5841	2.8043	0.4198	2.9618	5.6197	0.1259	0.1956	0.0000	0.6176	0.4378
4	1.2230	1.4284	0.3759	0.0000	5.2982	0.0000	0.1956	0.0000	0.0000	0.4378
5	1.1961	1.8581	0.4059	2.8145	5.4795	0.1259	0.1956	0.0000	0.6176	0.4378

续表

No	x_{21}	x_{26}	x_{28}	x_{29}	x_{32}	x_{48}	x_{51}	x_{56}	x_{59}	x_{77}
6	1.1040	4.7962	0.4453	4.7199	5.6838	0.1429	0.1956	0.0426	0.7018	0.4378
7	2.0224	2.0810	0.4195	2.8447	6.8442	0.1259	0.2067	0.0000	0.6399	0.4869
8	0.0000	4.0674	1.5173	3.2497	6.0225	0.1259	0.1956	0.0000	0.6176	0.4378
9	0.3216	3.0957	1.5034	3.1025	5.8823	0.1259	0.1956	0.0000	0.6176	0.4378
10	0.7934	3.0957	1.5034	3.1025	5.8823	0.1259	0.1956	0.0000	0.6176	0.4378
11	0.7269	2.4876	1.6649	2.0124	5.1643	0.1259	0.1956	0.0000	0.6176	0.4378
12	0.0000	3.9846	1.7352	3.1099	5.8912	0.1259	0.1956	0.0000	0.6176	0.4378
13	0.0000	5.4722	1.7791	3.3716	6.2128	0.1259	0.1956	0.0000	0.6176	0.4378
14	0.3216	4.4869	1.7652	3.2244	6.0726	0.1259	0.1956	0.0000	0.6176	0.4378
15	0.9173	1.5852	1.5177	1.9574	5.0884	0.1259	0.1956	0.0000	0.6176	0.4378
16	0.1903	3.0566	1.5880	3.0549	5.8153	0.1259	0.1956	0.0000	0.6176	0.4378
17	0.5119	3.5509	1.6180	3.1695	5.9967	0.1259	0.1956	0.0000	0.6176	0.4378
18	0.9837	3.5509	1.6180	3.1695	5.9967	0.1259	0.1956	0.0000	0.6176	0.4378
19	0.8916	6.5715	1.6573	5.1102	6.2010	0.1429	0.1956	0.0426	0.7018	0.4378

续表

No	x ₂₁	x ₂₆	x ₂₈	x ₂₉	x ₃₂	x ₄₈	x ₅₁	x ₅₆	x ₅₉	x ₇₇
20	1.8072	3.8018	1.6316	3.1996	7.3841	0.1259	0.2067	0.0000	0.6399	0.4869
21	0.8413	2.4876	1.6649	2.0124	5.1643	0.1259	0.1956	0.0000	0.6176	0.4378
22	0.1144	3.9846	1.7352	3.1099	5.8912	0.1259	0.1956	0.0000	0.6176	0.4378
23	0.1144	5.4722	1.7791	3.3716	6.2128	0.1259	0.1956	0.0000	0.6176	0.4378
24	0.4360	4.4869	1.7652	3.2244	6.0726	0.1259	0.1956	0.0000	0.6176	0.4378
25	0.2958	5.4722	1.7791	3.3716	6.2128	0.1259	0.1956	0.0000	0.6176	0.4378
26	0.9347	3.9846	1.7352	0.0000	5.8912	0.0000	0.1956	0.0000	0.0000	0.4378
27	0.8156	7.5164	1.8045	5.1688	6.2769	0.1429	0.1956	0.0426	0.7018	0.4378
28	0.7646	4.4869	3.1369	3.2244	7.9655	0.1914	0.6652	0.0000	0.6888	0.5646
29	1.0501	4.4869	1.7652	3.2244	9.4033	0.1259	0.2411	0.0000	0.7551	1.0221
30	1.0704	4.2523	3.0275	3.1738	7.7861	0.1674	0.6232	0.0000	0.6620	0.5249
31	1.4690	4.2523	1.7570	3.1738	9.1207	0.1259	0.2299	0.0000	0.7039	0.9457
32	0.8150	3.9846	1.8555	3.1099	5.8912	0.1545	0.2561	0.0000	0.6176	0.4378
33	1.6177	3.9846	1.7352	3.1099	6.4226	0.1259	0.2264	0.0000	0.7379	0.6258

续表

No	x_{21}	x_{26}	x_{28}	x_{29}	x_{32}	x_{48}	x_{51}	x_{56}	x_{59}	x_{77}
34	0.1364	5.4722	1.7791	3.3716	7.9039	0.1259	0.2264	0.0000	0.7379	0.6258
35	0.4238	2.6229	0.0000	4.4613	5.7009	0.0000	0.0000	0.1259	0.8132	0.4378
36	0.7454	3.0957	0.0000	4.6059	5.8823	0.0000	0.0000	0.1259	0.8132	0.4378
37	0.6052	4.0674	0.0000	4.7670	6.0225	0.0000	0.0000	0.1259	0.8132	0.4378
38	2.0588	3.3358	0.0000	4.6496	7.2607	0.0000	0.0000	0.1259	0.8466	0.4869
39	1.2172	3.0957	0.0000	4.6059	5.8823	0.0000	0.0000	0.1259	0.8132	0.4378
40	1.2381	1.5852	1.5177	1.9574	5.0884	0.1259	0.1956	0.0000	0.6176	0.4378
41	0.8327	3.5509	1.6180	3.1695	5.9967	0.1259	0.1956	0.0000	0.6176	0.4378
42	1.3314	3.0566	1.5880	0.0000	5.8153	0.0000	0.1956	0.0000	0.0000	0.4378
43	1.3045	3.5509	1.6180	3.1695	5.9967	0.1259	0.1956	0.0000	0.6176	0.4378
44	2.1618	3.8018	1.6316	3.1996	7.3841	0.1259	0.2067	0.0000	0.6399	0.4869
45	1.8995	6.2973	1.7188	5.0079	9.2149	0.1685	0.2299	0.0302	0.7800	0.9457
46	1.1559	9.5995	1.7664	7.0412	6.3388	0.1856	0.1956	0.0817	0.7600	0.4378
47	1.6743	3.7924	1.6481	3.1919	7.3595	0.1259	0.2180	0.0000	0.6349	0.4740
48	1.7019	3.0957	1.5034	3.1025	6.3644	0.1259	0.3159	0.0000	0.6871	0.5559

表 5-2B 58 肽组 25 个 MHDV 描述子

No	二肽	x_1	x_2	x_3	x_5	x_6	x_9	x_{14}	x_{15}
1	VW	0.2500	0.1800	2.9702	0.2291	0.1604	0.2288	5.0934	9.6804
2	IW	0.1111	1.4104	1.9999	0.1785	0.1331	0.1919	5.1833	11.1655
3	IY	0.1111	1.4082	1.9659	0.1785	0.1093	0.2083	3.9797	10.6536
4	AW	0.0000	0.1198	1.4301	0.2618	0.1388	0.1903	5.0934	9.5606
5	RW	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	7.6235	12.2461
6	VY	0.2500	0.1786	2.9321	0.2291	0.1340	0.2467	3.8903	9.1875
7	GW	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	5.2626	10.6816
8	VF	0.2500	0.2002	2.9105	0.2291	0.1340	0.2288	6.8187	6.5127
9	AY	0.0000	0.1208	1.4047	0.2618	0.1219	0.2013	3.8903	9.0668
10	IP	0.1111	1.4247	1.9084	0.1785	0.0000	0.1919	2.3522	4.4095
11	RP	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	4.8295	5.3421
12	AF	0.0000	0.1342	1.3912	0.2618	0.1219	0.1903	6.8187	6.3785
13	GY	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	4.0641	10.1478
14	AP	0.0000	0.1547	1.3617	0.2618	0.0000	0.1903	2.2500	2.8211

续表

No	二肽	x_1	x_2	x_3	x_5	x_6	x_9	x_{14}	x_{15}
44	LG	0.2500	0.5591	2.3902	0.1279	0.0845	0.1549	0.0686	3.7646
45	GD	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0686	3.6135
46	TG	0.0000	0.0431	1.3907	0.1146	0.0670	0.3833	0.0000	1.4970
47	EG	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.1117	4.3917
48	DG	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0686	3.7646
49	PG	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	2.4064	3.5352
50	LA	0.2930	0.5431	3.9234	0.1718	0.3463	0.4765	0.0000	2.3617
51	KA	0.0000	0.1105	1.4445	0.0569	0.2618	0.3216	3.6111	2.1604
52	RA	0.0000	0.0941	1.4576	0.0655	0.2783	0.3216	2.2500	2.4780
53	YA	0.0000	0.1208	1.4875	0.0439	0.2618	0.3325	3.8903	8.9113
54	AA	0.0431	0.0000	2.8062	0.3057	0.3837	0.5119	0.0000	0.0000
55	FR	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	9.3808	8.9545
56	HL	0.2500	0.5903	2.5252	0.0436	0.1753	0.2330	0.7652	6.6917
57	DA	0.0000	0.0431	1.4740	0.0439	0.2618	0.3664	0.0000	2.3617
58	EA	0.0000	0.0726	1.4660	0.0439	0.2618	0.3555	1.0000	3.0454

表 5-2B 58 肽组 25 个 MHDV 描述子(续)

No	x_{17}	x_{18}	x_{19}	x_{21}	x_{26}	x_{28}	x_{29}	x_{30}	x_{32}
1	0.1215	2.2722	0.0000	0.7013	7.5164	1.8045	5.1688	0.0000	6.2769
2	0.2360	2.3523	0.0000	0.8156	7.5164	1.8045	5.1688	0.0000	6.2769
3	0.2371	0.5469	0.0000	1.7293	4.7404	1.7788	3.2546	0.0000	7.4620
4	0.1215	2.2722	0.0000	0.7013	6.0864	1.5428	5.0300	0.0000	6.0866
5	1.1213	4.0751	0.0000	1.0834	6.2132	4.2691	6.3322	0.0000	6.1306
6	0.1226	0.4799	0.0000	1.6060	4.7404	1.7788	3.2546	0.0000	7.4620
7	1.2190	2.5823	0.0000	1.1040	4.7962	0.4453	4.7199	0.0000	5.6838
8	0.1362	0.5100	0.0000	0.7934	4.4869	1.7652	3.2244	0.0000	6.0726
9	0.1226	0.4799	0.0000	1.6060	3.3358	1.5170	3.1326	0.0000	7.2607
10	0.2726	0.0000	1.6880	0.9347	3.9846	1.7352	0.0000	3.1099	5.8912
11	1.1467	1.5287	1.8522	1.2025	2.7229	4.1542	1.2405	3.0097	5.7449
12	0.1362	0.5100	0.0000	0.7934	3.0957	1.5034	3.1025	0.0000	5.8823
13	1.2201	0.7678	0.0000	2.0224	2.0810	0.4195	2.8447	0.0000	6.8442
14	0.1581	0.0000	1.6210	0.8203	2.6229	1.4734	0.0000	2.9879	5.7009

续表

No	x_{17}	x_{18}	x_{19}	x_{21}	x_{26}	x_{28}	x_{29}	x_{30}	x_{32}
15	1.1430	2.2779	0.0000	1.1755	3.2063	4.2020	4.3778	0.0000	5.9263
16	0.1581	0.0000	1.6210	0.8203	3.9846	1.7352	0.0000	3.1099	5.8912
17	1.2556	0.0000	1.9090	1.2230	1.4284	0.3759	0.0000	2.7000	5.2982
18	1.2337	0.7979	0.0000	1.1961	1.8581	0.4059	2.8145	0.0000	5.4795
19	0.2507	0.5770	0.0000	0.9077	4.4869	1.7652	3.2244	0.0000	6.0726
20	0.0703	1.0975	0.0000	0.7269	2.4876	1.6649	2.0124	0.0000	5.1643
21	0.1848	1.1645	0.0000	0.8413	2.4876	1.6649	2.0124	0.0000	5.1643
22	1.1275	0.4025	0.0000	0.5841	2.8043	0.4198	2.9618	0.0000	5.6197
23	1.1714	0.6642	0.0000	0.9057	1.4284	0.3759	2.7000	0.0000	5.2982
24	1.0975	0.2879	0.0000	0.4028	1.4284	0.3759	2.7000	0.0000	5.2982
25	0.5501	1.3671	0.0000	2.0824	1.7987	1.5479	1.9799	0.0000	6.4285
26	1.1414	0.5497	0.0000	0.7243	1.8581	0.4059	2.8145	0.0000	5.4795
27	0.0703	1.0975	0.0000	0.7269	1.1945	1.4032	1.8904	0.0000	4.9740
28	1.1806	4.8587	0.0000	0.9275	1.8581	0.4059	4.5675	0.0000	5.4795

续表

No	x_{17}	x_{18}	x_{19}	x_{21}	x_{26}	x_{28}	x_{29}	x_{30}	x_{32}
29	1.6985	2.2238	0.0000	1.0222	1.5206	3.0268	3.9613	0.0000	5.3595
30	2.1056	1.3578	0.0000	1.1655	1.1945	1.4851	1.8904	0.0000	4.9740
31	0.5803	1.3896	0.0000	1.2381	1.5852	1.5177	1.9574	0.0000	5.0884
32	1.1414	0.5497	0.0000	1.9285	1.4284	0.3759	2.7000	0.0000	5.7548
33	1.0975	0.2879	0.0000	0.4028	2.8043	0.4198	2.9618	0.0000	5.6197
34	0.4466	1.2864	0.0000	1.0317	1.1945	1.4032	1.8904	0.0000	4.9740
35	2.7641	0.7689	0.0000	1.1033	1.4284	0.4667	2.7000	0.0000	5.2982
36	1.1714	0.6642	0.0000	1.7756	1.6451	0.3977	2.7639	0.0000	8.4489
37	1.0975	0.2879	0.0000	0.4471	2.8043	0.4198	2.9618	0.0000	7.2362
38	0.5237	3.1965	0.0000	1.1816	4.4869	1.6268	3.8136	0.0000	5.2262
39	0.4404	5.4936	0.0000	1.0544	1.5852	1.5177	3.6741	0.0000	5.0884
40	1.5695	0.6642	0.0000	1.3682	1.6451	1.6299	2.7639	0.0000	7.1529
41	1.1678	1.3854	0.0000	1.1297	0.1259	0.3057	1.6025	0.0000	4.5712
42	0.8447	1.2864	0.0000	1.4941	1.3896	2.6859	1.9327	0.0000	6.7752

续表

No	x_{17}	x_{18}	x_{19}	x_{21}	x_{26}	x_{28}	x_{29}	x_{30}	x_{32}
43	0.3320	1.2195	0.0000	2.1214	1.1945	1.4032	1.8904	0.0000	5.3897
44	0.3320	1.2195	0.0000	0.9173	1.5852	1.5177	1.9574	0.0000	5.0884
45	1.1414	0.5497	0.0000	1.3666	1.8581	0.4059	2.8145	0.0000	8.7033
46	0.0703	1.0975	0.0000	0.7713	2.4876	1.6649	2.0124	0.0000	6.7398
47	0.4466	1.2864	0.0000	1.9016	1.3896	1.4671	1.9327	0.0000	8.0576
48	0.3320	1.2195	0.0000	1.5595	1.5852	1.5177	1.9574	0.0000	8.2656
49	0.0000	3.0480	0.0000	1.1507	1.1945	0.0000	3.2936	0.0000	4.9740
50	0.2618	0.1219	0.0000	0.1903	3.0566	1.5880	3.0549	0.0000	5.8153
51	2.0188	0.2603	0.0000	0.4385	2.6229	1.5719	2.9879	0.0000	5.7009
52	0.9186	1.7050	0.0000	0.3822	2.7229	4.1542	4.2503	0.0000	5.7449
53	0.4799	0.2696	0.0000	1.3418	3.2874	1.6181	3.0774	0.0000	7.1691
54	0.0000	0.0000	0.0000	0.0000	2.6229	1.4734	2.9879	0.0000	5.7009
55	1.1723	2.2691	0.0000	1.1306	3.1762	4.2835	4.3510	0.0000	5.8766
56	0.4140	4.6441	0.0000	0.6490	3.5509	1.6180	4.9685	0.0000	5.9967

续表

No	x_{17}	x_{18}	x_{19}	x_{21}	x_{26}	x_{28}	x_{29}	x_{30}	x_{32}
57	0.2618	0.1219	0.0000	0.7705	3.0566	1.5880	3.0549	0.0000	9.0546
58	0.3763	0.1889	0.0000	1.1298	2.8476	1.5374	3.0302	0.0000	8.8294

表 5-2B 58 肽组 25 个 MHDV 描述子(续)

No	x_{47}	x_{48}	x_{49}	x_{51}	x_{56}	x_{59}	x_{66}	x_{77}
1	0.0000	0.1429	0.0000	0.1956	0.0426	0.7018	0.0000	0.4378
2	0.0000	0.1429	0.0000	0.1956	0.0426	0.7018	0.0000	0.4378
3	0.0000	0.1259	0.0000	0.2067	0.0000	0.6399	0.0000	0.4869
4	0.0000	0.1429	0.0000	0.1956	0.0426	0.7018	0.0000	0.4378
5	0.3528	0.8147	0.0000	0.2666	0.0810	0.7589	0.0000	0.4378
6	0.0000	0.1259	0.0000	0.2067	0.0000	0.6399	0.0000	0.4869
7	0.0000	0.1429	0.0000	0.1956	0.0426	0.7018	0.0000	0.4378
8	0.0000	0.1259	0.0000	0.1956	0.0000	0.6176	0.0000	0.4378
9	0.0000	0.1259	0.0000	0.2067	0.0000	0.6399	0.0000	0.4869

续表

No	x_{47}	x_{48}	x_{49}	x_{51}	x_{56}	x_{59}	x_{66}	x_{77}
10	0.0000	0.0000	0.1259	0.1956	0.0000	0.0000	0.6176	0.4378
11	0.3528	0.6250	0.1599	0.2666	0.0000	0.0571	0.6176	0.4378
12	0.0000	0.1259	0.0000	0.1956	0.0000	0.6176	0.0000	0.4378
13	0.0000	0.1259	0.0000	0.2067	0.0000	0.6399	0.0000	0.4869
14	0.0000	0.0000	0.1259	0.1956	0.0000	0.0000	0.6176	0.4378
15	0.3528	0.7849	0.0000	0.2666	0.0295	0.6747	0.0000	0.4378
16	0.0000	0.0000	0.1259	0.1956	0.0000	0.0000	0.6176	0.4378
17	0.0000	0.0000	0.1259	0.1956	0.0000	0.0000	0.6176	0.4378
18	0.0000	0.1259	0.0000	0.1956	0.0000	0.6176	0.0000	0.4378
19	0.0000	0.1259	0.0000	0.1956	0.0000	0.6176	0.0000	0.4378
20	0.0000	0.1259	0.0000	0.1956	0.0000	0.6176	0.0000	0.4378
21	0.0000	0.1259	0.0000	0.1956	0.0000	0.6176	0.0000	0.4378
22	0.0000	0.1259	0.0000	0.1956	0.0000	0.6176	0.0000	0.4378
23	0.0000	0.1259	0.0000	0.1956	0.0000	0.6176	0.0000	0.4378

续表

No	x_{47}	x_{48}	x_{49}	x_{51}	x_{56}	x_{59}	x_{66}	x_{77}
24	0.0000	0.1259	0.0000	0.1956	0.0000	0.6176	0.0000	0.4378
25	0.0000	0.1259	0.0000	0.2180	0.0000	0.6349	0.0000	0.4740
26	0.0000	0.1259	0.0000	0.1956	0.0000	0.6176	0.0000	0.4378
27	0.0000	0.1259	0.0000	0.1956	0.0000	0.6176	0.0000	0.4378
28	0.0000	0.1652	0.0000	0.1956	0.3947	0.8215	0.0000	0.4378
29	0.3309	0.7699	0.0000	0.2919	0.0415	0.6999	0.0000	0.4378
30	0.0286	0.1474	0.0000	0.2391	0.0000	0.6176	0.0000	0.4378
31	0.0000	0.1259	0.0000	0.1956	0.0000	0.6176	0.0000	0.4378
32	0.0000	0.1259	0.0000	0.2264	0.0000	0.7379	0.0000	0.6258
33	0.0000	0.1259	0.0000	0.1956	0.0000	0.6176	0.0000	0.4378
34	0.0000	0.1259	0.0000	0.1956	0.0000	0.6176	0.0000	0.4378
35	0.0131	0.1545	0.0000	0.2561	0.0000	0.6176	0.0000	0.4378
36	0.0000	0.1259	0.0000	0.2299	0.0000	0.7039	0.0000	0.9457
37	0.0000	0.1259	0.0000	0.2264	0.0000	0.7379	0.0000	0.6258

续表

No	x_{47}	x_{48}	x_{49}	x_{51}	x_{56}	x_{59}	x_{66}	x_{77}
38	0.0000	0.1685	0.0000	0.1956	0.0302	0.6758	0.0000	0.4378
39	0.0000	0.2347	0.0000	0.1956	0.3596	0.7552	0.0000	0.4378
40	0.0167	0.1674	0.0000	0.6232	0.0000	0.6620	0.0000	0.5249
41	0.0000	0.1259	0.0000	0.1956	0.0000	0.6176	0.0000	0.4378
42	0.0415	0.1555	0.0000	0.6248	0.0000	0.6488	0.0000	0.4978
43	0.0000	0.1259	0.0000	0.3159	0.0000	0.6871	0.0000	0.5559
44	0.0000	0.1259	0.0000	0.1956	0.0000	0.6176	0.0000	0.4378
45	0.0000	0.1259	0.0000	0.2411	0.0000	0.7551	0.0000	1.0221
46	0.0000	0.1259	0.0000	0.3159	0.0000	0.6871	0.0000	0.5559
47	0.0000	0.1259	0.0000	0.2819	0.0000	0.6786	0.0000	0.8931
48	0.0000	0.1259	0.0000	0.3332	0.0000	0.7072	0.0000	0.9384
49	0.0000	0.0000	0.0000	0.0000	0.1259	0.8132	0.0000	0.4378
50	0.0000	0.1259	0.0000	0.1956	0.0000	0.6176	0.0000	0.4378
51	0.0286	0.1474	0.0000	0.2391	0.0000	0.6176	0.0000	0.4378

续表

No	x_{47}	x_{48}	x_{49}	x_{51}	x_{56}	x_{59}	x_{66}	x_{77}
52	0.3528	0.7849	0.0000	0.2666	0.0295	0.6747	0.0000	0.4378
53	0.0000	0.1259	0.0000	0.2180	0.0000	0.6349	0.0000	0.4740
54	0.0000	0.1259	0.0000	0.1956	0.0000	0.6176	0.0000	0.4378
55	0.3309	0.7699	0.0000	0.2919	0.0415	0.6999	0.0000	0.4378
56	0.0000	0.2347	0.0000	0.1956	0.3596	0.7552	0.0000	0.4378
57	0.0000	0.1259	0.0000	0.3332	0.0000	0.7072	0.0000	0.9384
58	0.0000	0.1259	0.0000	0.2819	0.0000	0.6786	0.0000	0.8931

表 5-3 两组二肽分子的 MHDV 模式空间距离分布

No	48 肽组		58 肽组	
	距离范围	频率/%	距离范围	频率/%
1	0.256~1.5	0.53	0.127~1	0.12
2	1.5~4.5	22.07	1~3	6.96
3	4.5~7.5	26.42	3~5	24.44
4	7.5~10.5	23.05	5~7	22.69
5	10.5~13.5	16.49	7~9	22.02
6	13.5~16.5	5.23	9~11	14.34
7	16.5~19.5	3.37	11~13	7.14
8	19.5~22.5	2.22	13~15	1.81
9	22.5~25.5	0.62	15~17	0.48

5.1.2 最佳主成分数的选择

由于二肽分子 MHDV 描述子较多, 建立二肽分子生物活性与 MHDV 描述子之间的定量关系时, 宜采用主成分回归 (PCR) 技术。在进行 PCR 分析时, 最佳主成分数的确定常常是最重要的步骤之一, 这个主成分选取的好坏直接影响所建模型的稳定性、可估计性和可预测性。研究发现, 在建立 PCR 模型时, 估计生物活性与实验生物活性之间的相关系数 (r) 和均方根误差 ($RMSE$) 的变化是很有规律的。一般说来, $RMSE$ 随着主成分数 (PC) 的增加而降低 相反 r 则随主成分数 PC 的增加而升高。但是, 在验证模型稳定性或预测能力所进行的 LOO 交互检验过程中, PC 的变化与 QSAR 统计结果之间的变化关系不存在明显的规律性。其中 LOO 校验均方根误差 $RMSV$ 随 PC 的变化曲线中可能存在着多个极小值, 而 LOO 检验相关系数 q 则可能存在多个极大值。表

5-4分别给出了 48 肽组和 58 肽组在建模与 LOO 检验过程中不同主成分 PC 对模型统计量的影响。图 5-3 则展示了 48 肽组与 58 肽组 QSAR 模型 LOO 交互检验预测过程中的均方根误差随 PC 的变化曲线。

表 5-4 两组二肽不同主成分对 QSAR 模型的影响

PC	$n = 48, m = 20$				$n = 58, m = 33$			
	RMSE	r	RMSV	q	RMSE	r	RMSV	q
1	0.568	0.4108	0.584	0.3499	0.979	0.1645	0.987	0.1057
2	0.558	0.4426	0.583	0.3536	0.953	0.2788	1.019	0.0455
3	0.548	0.4760	0.612	0.2737	0.673	0.7352	0.711	0.6985
4	0.465	0.6647	0.516	0.5597	0.661	0.7460	0.729	0.6808
5	0.339	0.8384	0.394	0.7750	0.552	0.8309	0.607	0.7917
6	0.335	0.8426	0.394	0.7766	0.519	0.8526	0.585	0.8081
7	0.283	0.8907	0.368	0.8093	0.492	0.8687	0.571	0.8181
8	0.235	0.9264	0.300	0.8782	0.481	0.8747	0.552	0.8315
9	0.211	0.9408	0.285	0.8903	0.469	0.8813	0.549	0.8338
10	0.197	0.9487	0.257	0.9122	0.467	0.8823	0.557	0.8287
11	0.179	0.9578	0.255	0.9139	0.466	0.8828	0.573	0.8180
12	0.169	0.9625	0.235	0.9273	0.464	0.8839	0.612	0.7905
13	0.158	0.9671	0.226	0.9328	0.367	0.9293	0.474	0.8802
14	0.158	0.9672	0.230	0.9306	0.357	0.9332	0.469	0.8826
15	0.153	0.9695	0.221	0.9363	0.355	0.9338	0.475	0.8793
16	0.150	0.9705	0.232	0.9301	0.355	0.9338	0.491	0.8718
17	0.140	0.9742	0.214	0.9405	0.353	0.9345	0.491	0.8725
18	0.138	0.9752	0.220	0.9371	0.345	0.9376	0.505	0.8650

续表

PC	$n = 48, m = 20$				$n = 58, m = 33$			
	RMSE	r	RMSV	q	RMSE	r	RMSV	q
19	0.137	0.975 6	0.232	0.929 9	0.338	0.940 2	0.502	0.866 8
20	0.131	0.977 5	0.220	0.938 0	0.330	0.943 2	0.493	0.871 3

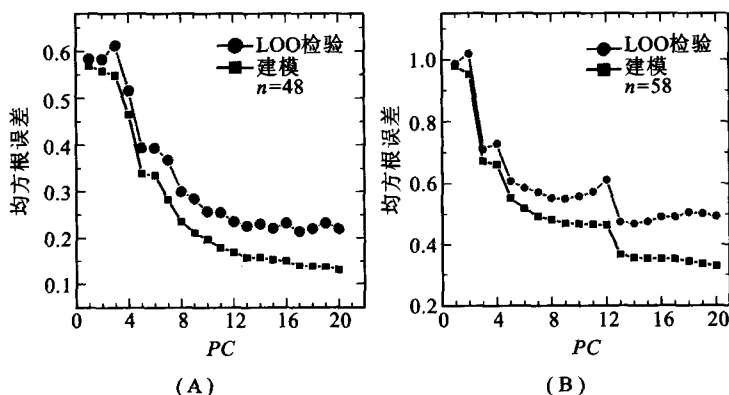


图 5-3 48 肽组 (A) 及 58 肽组 (B) 建模与 LOO 检验过程中均方根误差随主成分的变化曲线

由表 5-4 和图 5-3 可知, 均方根误差 (RMSE) 在建模过程中明显随 PC 增加而降低; r 明显随 PC 增加而升高。然而, 在 LOO 交互检验预测过程中, 很难找到 RMSV 和 q 随 PC 增加而发生变化的规律。但是, r 的最大值与 RMSE 的最小值所对应的主成分是一致的。如 48 肽组, 当 $PC = 17$ 时, RMSV 有最小值 (0.214) 此时 q 有最大值 0.9405 同样对于 58 肽组当 $PC = 14$ 时, RMSV 有最小值 (0.69), 而 q 有最大值 0.8826。因此, 主成分的判别可以根据 RMSV 或者 q 进行判别与选择。本章采用 RMSV 为标准选择最佳主成分数。

对于 48 肽组, PC 从 12 变化到 $PC = 20$ 时, $RMSV$ 具有相对稳定的低值, 其变化很小, 因此合适的主成分数应为 $PC = 12$ 。对于 58 肽组, PC 从 13 变化到 20 的范围内, $RMSV$ 值趋于稳定, 因此 $PC = 13$ 是合适的。最佳主成分模型分别为:

对于 48 肽组:

$$n = 48, m = 20, PC = 12, r = 0.9625, RMSE = 0.169 \quad (\text{建模})$$

$$n = 48, m = 20, PC = 12, q = 0.9273, RMSV = 0.235 \quad (\text{LOO 检验})$$

对于 58 肽组:

$$n = 58, m = 25, PC = 13, r = 0.9293, RMSE = 0.367 \quad (\text{建模})$$

$$n = 58, m = 25, PC = 13, q = 0.8802, RMSV = 0.474 \quad (\text{LOO 检验})$$

由上述主成分回归分析结果可知, 原始 MHDV 描述子变量中, 由于变量之间可能存在的线性相关问题, 并不是每一个 MHDV 描述子变量都是独立的, 因而有些变量重要, 而另一些变量相对的不那么重要。换句话说, 48 肽组中的 20 个非零 MHDV 描述子中可能只有少数 MHDV 描述子对生物活性有重要影响。若应用全部的 MHDV 变量建模, 其模型可能是不稳定的或不具预测能力。例如, 应用 48 肽组的 20 个 MHDV 描述子和 58 肽组的 25 个 MHDV 描述子, 建立多元线性回归模型, 有:

对于 48 肽组:

$$n = 48, m = 20, r = 0.9775, RMSE = 0.131 \quad (\text{建模})$$

$$n = 48, m = 20, q = 0.9320, RMSV = 0.232 \quad (\text{LOO 检验})$$

对于 58 肽组:

$$n = 58, m = 25, r = 0.9502, RMSE = 0.309 \quad (\text{建模})$$

$$n = 58, m = 25, q = 0.1105, RMSV = 39.799 \quad (\text{LOO 检验})$$

上述结果表明, 对于 48 肽组, 全部 MHDV 变量的 MLR 模型与 LOO 检验结果与主成分回归结果没有显著性差异。但对于 58 肽组, 全部变量的 MLR 模型的 LOO 检验结果由于第 11 号二肽 (RP) 的 LOO 预测误差极大, 造成 q 很低与 $RMSV$ 很高。因此, 对于变量数较多的体系, 在建立 QSAR 模型时, 要么应用基于主成分的方法, 要么先对变量进行优化选择。

5.1.3 主成分回归模型与似回归 (MLR-like) 模型

前面已经说明, 最佳主成分选择对 PCR 模型结果的重要影响。然而, 主成分回归模型不是一个显式模型, 而是一个关于主成分的隐含模型。如果直接使用这个隐含模型进行估计与预测, 就不能直观地观察生物活性与 MHDV 原始描述子之间的变化规律。因此, 常常要进行变换, 以构建习用的多变量线性模型即似回归模型, 其变换原理参见第三章关于 PCR 的有关内容。所谓似回归模型, 只是形式上类似于多元线性回归模型, 但模型中的原始变量不是完全独立的, 可能是相关的。这个模型在变量数大于样本数时用一般多元回归方法是不能构建的。有了似回归模型, 对已知或未知生物活性化合物进行估计或预测时, 只需将化合物的有关描述子变量代入似回归模型即可, 而不必再进行主成分分析, 使用非常方便。

对于两组二肽, 主成分模型与似回归模型总结如下:

对于 48 肽组, 主成分回归模型统计量为:

$$n = 48, m = 20, PC = 12, r = 0.9625, RMSE = 0.169 \quad (\text{建模})$$

$$n = 48, m = 20, PC = 12, q = 0.9273, RMSV = 0.235 \quad (\text{LOO 检验})$$

相应的似回归模型为:

$$\begin{aligned}
A_{48} = & 1.7656 + 1.2721x_1 + 0.3664x_2 + 0.0534x_3 - \\
& 1.5384x_5 - 0.5509x_6 - 0.1927x_9 + \\
& 0.0950x_{14} + 0.0539x_{15} - 0.1031x_{17} - \\
& 0.0161x_{18} - 0.0717x_{21} + 0.0682x_{26} + \\
& 0.1817x_{28} + 0.0201x_{29} - 0.0415x_{32} - \\
& 0.9638x_{48} - 0.8623x_{51} + 0.0857x_{56} - \\
& 0.2015x_{59} - 0.9083x_{77}
\end{aligned} \tag{5-1}$$

对于 58 肽组，主成分回归模型统计量为：

$$n = 58, m = 25, PC = 13, r = 0.9293, RMSE = 0.367 \quad (\text{建模})$$

$$n = 58, m = 25, PC = 13, q = 0.8802, RMSV = 0.474 \quad (\text{LOO检验})$$

相应的似回归模型为：

$$\begin{aligned}
A_{58} = & 1.0688 - 0.9230x_1 + 0.0440x_2 + 0.0590x_3 + \\
& 3.9953x_5 + 0.3786x_6 + 0.2088x_9 - \\
& 0.0360x_{14} + 0.0835x_{15} + 0.3088x_{17} - \\
& 0.0579x_{18} + 0.1339x_{21} + 0.4690x_{22} + \\
& 0.2554x_{26} - 0.1023x_{28} + 0.2315x_{29} + \\
& 0.0768x_{32} + 0.0391x_{47} + 0.6603x_{48} + \\
& 0.3501x_{51} + 1.8167x_{56} - 0.7120x_{59} - \\
& 1.2363x_{77} - 0.1660x_{59} + 0.3835x_{77} - \\
& 1.4194x_{59}
\end{aligned} \tag{5-2}$$

由模型 (5-1) 式估计的 48 个肽的苦味活性值 (BT_{EST}) 及 LOO 交互检验预测的活性值 (BT_{LOO}) 与原实验活性值 (BT_{OBS}) 均列入表 5-5A, 估计值总体相关情况见图 5-4A。由模型 (5-2) 式估计的 58 个肽的半数抑制浓度活性 ($pIC_{50,EST}$) 及 LOO 交互检验预测的活性值 ($pIC_{50,LOO}$) 与原实验活性值 ($pIC_{50,OBS}$) 均列入表 5-5B, 估计值总体相关情况见图 5-4B。

表 5-5A 48 个二肽苦味活性的实验值(BT_{OBS})、估计值(BT_{EST})与 LOO 检验预测值(BT_{LOO})

No	BT_{OBS}	BT_{EST}	BT_{LOO}	No	BT_{OBS}	BT_{EST}	BT_{LOO}	No	BT_{OBS}	BT_{EST}	BT_{LOO}
1	1.130	1.183	1.196	17	2.350	2.408	2.413	33	1.490	1.479	1.529
2	1.680	1.477	1.439	18	2.750	2.740	2.733	34	1.490	1.490	1.494
3	1.700	1.524	1.482	19	3.400	2.927	2.854	35	1.320	1.407	1.459
4	1.350	1.370	1.384	20	2.460	2.444	2.441	36	2.220	2.203	2.197
5	1.800	1.824	1.825	21	1.680	1.666	1.665	37	2.330	2.257	2.232
6	1.890	2.021	2.056	22	1.680	1.626	1.618	38	1.800	2.225	2.462
7	1.770	1.537	1.459	23	2.050	2.152	2.161	39	2.800	2.511	2.387
8	1.160	1.153	1.150	24	2.260	2.447	2.470	40	1.770	2.065	2.170
9	1.700	1.446	1.420	25	2.260	2.500	2.573	41	2.870	2.824	2.814
10	1.720	1.786	1.804	26	2.400	2.362	2.333	42	2.700	2.689	2.681
11	1.190	1.283	1.319	27	3.050	2.962	2.939	43	3.100	3.163	3.175
12	1.160	1.241	1.264	28	1.490	1.376	1.242	44	3.130	2.871	2.786
13	1.710	1.764	1.786	29	1.370	1.115	0.937	45	1.560	1.909	2.133
14	2.000	2.058	2.070	30	1.490	1.570	1.633	46	3.600	3.567	3.529
15	1.720	1.633	1.611	31	1.370	1.316	1.296	47	2.400	2.535	2.568
16	1.720	1.590	1.577	32	1.650	1.962	2.105	48	1.490	1.524	1.420

表 5-5B 58 个二肽的 $pI_{C_{90}}$ 的实验值($pI_{C_{OBS}}$)、估计值($pI_{C_{EST}}$)与 LOO 检验预测值($pI_{C_{LOO}}$)

No	$pI_{C_{OBS}}$	$pI_{C_{EST}}$	$pI_{C_{LOO}}$	No	$pI_{C_{OBS}}$	$pI_{C_{EST}}$	$pI_{C_{LOO}}$	No	$pI_{C_{OBS}}$	$pI_{C_{EST}}$	$pI_{C_{LOO}}$
1	5.800	5.202	4.975	21	2.920	2.865	2.849	41	2.140	1.804	1.766
2	5.700	5.312	5.145	22	2.920	3.055	3.089	42	2.130	1.898	1.733
3	5.430	4.715	4.444	23	2.850	2.475	2.446	43	2.070	2.046	2.007
4	5.000	5.066	5.064	24	2.700	2.445	2.403	44	2.060	2.436	2.504
5	4.800	4.772	4.760	25	2.700	2.965	3.026	45	2.040	2.051	2.047
6	4.660	4.600	4.577	26	2.600	2.717	2.739	46	2.000	2.271	2.351
7	4.520	4.130	4.028	27	2.600	2.673	2.679	47	2.000	1.781	1.752
8	4.280	3.845	3.746	28	2.510	2.386	2.309	48	1.850	1.527	1.463
9	4.060	4.477	4.601	29	2.490	2.863	2.966	49	1.770	1.991	2.115
10	3.890	3.941	3.969	30	2.490	2.299	2.294	50	3.510	3.065	2.974
11	3.740	3.590	3.497	31	2.430	2.204	2.141	51	3.420	3.017	2.939
12	3.720	3.726	3.729	32	2.420	2.555	2.499	52	3.340	3.110	3.018
13	3.680	3.565	3.527	33	2.340	2.789	2.890	53	3.340	3.649	3.770
14	3.640	3.723	3.752	34	2.320	1.991	1.959	54	3.210	3.283	3.291
15	3.640	3.435	3.389	35	2.270	2.845	3.014	55	3.040	3.341	3.439
16	3.380	3.812	3.995	36	2.270	2.313	2.319	56	2.490	2.874	3.117
17	3.350	2.954	2.803	37	2.240	2.564	2.612	57	2.420	2.178	2.101
18	3.200	2.812	2.768	38	2.230	3.525	3.868	58	2.000	2.440	2.445
19	3.030	3.955	4.255	39	2.200	1.863	1.719				
20	2.960	2.758	2.712	40	2.150	2.418	2.563				

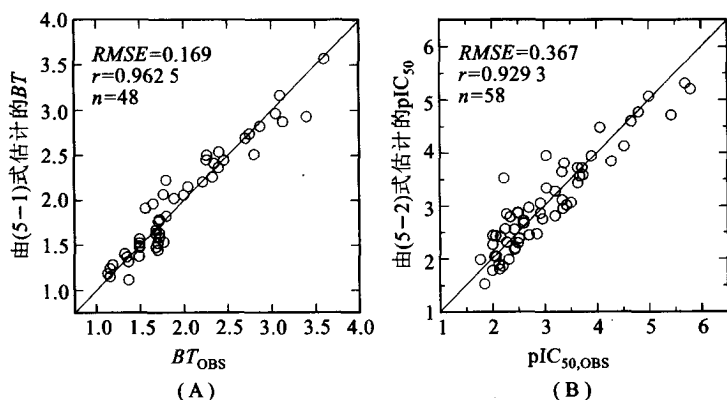


图 5-4 48 肽组 (A) 和 58 肽组 (B) 实验活性与 PCR 模型估计值相关图

5.1.4 偏最小二乘回归模型与似回归模型

前面应用主成分回归技术研究了两组二肽分子全息距离矢量描述子与其生物活性的相关关系，结果表明 MHDV 描述子能很好地描述这些二肽分子的结构信息，所构建的模型具有良好的估计能力和稳定性。然而，从各自最佳主成分数看，对于 48 个分子的二肽组和 58 个分子的二肽组来说， $PC=12$ 或 13 有悖于一般的经验准则，有点偏大。通过对大量化合物体系的 QSAR 研究结果进行分析可以得出，这是由于 PCR 方法在构建模型过程中只对分子描述子数据矩阵进行主成分提取，其主成分并不依赖于二肽分子的生物活性，从而使得最佳主成分偏大。本节将应用偏最小二乘回归 (PLSR) 方法重新对两组二肽的 MHDV 描述子与相应生物活性进行分析，建立有关 QSAR 模型并进行检验。

对于 48 肽组，PLSR 分析结果表明，PLSR 模型的估计均方根误差 RMSE 与 LOO 检验均方根误差 RMSV 随主成分的变化曲线如图 5-5A 所示，估计与 LOO 检验的相关系数与均方根误差数据见表 5-6。由图 5-5A 与表 5-6 可知，最佳主成分数应

为 5 虽然最佳主成分数 PC 比 PCR 模型的最佳主成分数 $PC = 12$ 小了很多,但有关模型统计量结果与 PCR 模型并没有显著性差异, PC 的判别也比 PCR 方法更容易,这也表明 PLSR 模型更有说服力。与 PLSR 模型相关的几个重要统计量如下:

$$n = 48, m = 20, PC = 5, r = 0.9621, RMSE = 0.170 \quad (\text{建模})$$

$$n = 48, m = 20, PC = 5, q = 0.9285, RMSV = 0.233$$

(LOO 检验)

对于 58 肽组,PLSR 分析结果表明,PLSR 模型的估计均方根误差 $RMSE$ 与 LOO 检验均方根误差 $RMSV$ 随主成分的变化曲线如图 5-5B 所示,估计与 LOO 检验的相关系数与均方根误差数据见表 5-6。由图 5-5B 与表 5-6 可知最佳主成分数 PC 应为 7 这比 PCR 模型的最佳主成分 $PC = 13$ 也小了很多,但有关统计量结果与 PCR 模型也没有显著性差异,这表明 PLSR 模型较 PCR 模型更优。与 PLSR 模型相关的几个重要统计量如下:

$$n = 58, m = 25, PC = 7, r = 0.9340, RMSE = 0.355 \quad (\text{建模})$$

$$n = 58, m = 25, PC = 7, q = 0.8758, RMSV = 0.483 \quad (\text{LOO 检验})$$

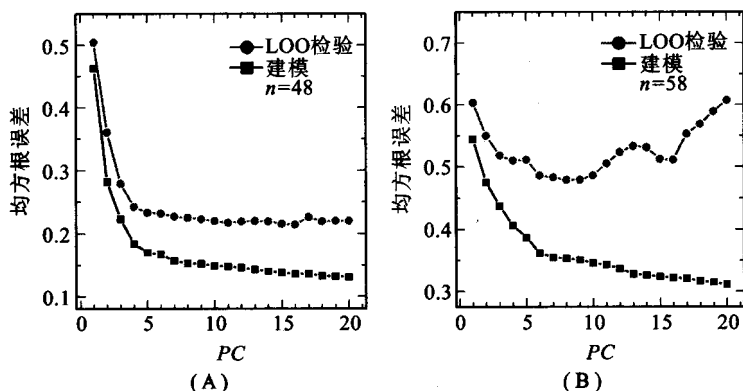


图 5-5 48 肽组 (A) 与 58 肽组 (B) PLSR 模型中均方根误差随主成分的变化曲线

表 5-6 两组二肽 PLS 模型不同主成分数对相关系数与均方根误差的影响

PC	48 肽组 ($m = 20$)				58 肽组 ($m = 25$)			
	r	RMSE	q	RMSV	r	RMSE	q	RMSV
1	0.6685	0.463	0.5894	0.504	0.8362	0.544	0.7953	0.603
2	0.8912	0.282	0.8170	0.360	0.8783	0.475	0.8333	0.550
3	0.9335	0.223	0.8938	0.279	0.8979	0.437	0.8535	0.518
4	0.9551	0.184	0.9216	0.242	0.9126	0.406	0.8588	0.510
5	0.9621	0.170	0.9285	0.233	0.9209	0.387	0.8598	0.511
6	0.9635	0.167	0.9287	0.232	0.9312	0.362	0.8735	0.486
7	0.9675	0.157	0.9323	0.227	0.9340	0.355	0.8758	0.483
8	0.9692	0.153	0.9338	0.225	0.9345	0.353	0.8772	0.479
9	0.9699	0.152	0.9344	0.223	0.9353	0.351	0.8780	0.479
10	0.9709	0.149	0.9364	0.220	0.9375	0.346	0.8751	0.486
11	0.9713	0.148	0.9380	0.217	0.9383	0.343	0.8647	0.505
12	0.9721	0.146	0.9375	0.219	0.9405	0.337	0.8549	0.523
13	0.9733	0.143	0.9374	0.220	0.9437	0.328	0.8499	0.533
14	0.9742	0.140	0.9382	0.219	0.9445	0.326	0.8513	0.531
15	0.9750	0.138	0.9397	0.215	0.9452	0.324	0.8624	0.512
16	0.9758	0.136	0.9399	0.214	0.9458	0.322	0.8625	0.511
17	0.9759	0.136	0.9342	0.226	0.9464	0.321	0.8397	0.553
18	0.9769	0.133	0.9370	0.219	0.9478	0.317	0.8295	0.569
19	0.9774	0.132	0.9373	0.220	0.9482	0.315	0.8193	0.589
20	0.9775	0.131	0.9380	0.220	0.9497	0.311	0.8102	0.607

偏最小二乘回归模型也是一个关于主成分的隐含模型，不是关于原始 MHDV 描述子变量的显式模型。经过适当变换（参见第三章）也可以将 PLSR 模型变换为关于原始描述子变量的所谓似回归模型。有了似回归模型，就可利用化合物的原始描述子变量进行生物活性的估计与预测，使用非常方便。相对于 PLSR 模

型的似回归模型结果如下：

对于 48 肽组，似回归模型 ($PC = 5$) 如下：

$$\begin{aligned}
 A_{48} = & 1.9080 + 1.1387x_1 + 0.3712x_2 + 0.0556x_3 - \\
 & 1.2110x_5 - 0.5286x_6 - 0.1712x_9 + 0.0799x_{14} + \\
 & 0.0661x_{15} - 0.1457x_{17} - 0.0328x_{18} - 0.0072x_{21} + \\
 & 0.0755x_{26} + 0.0828x_{28} + 0.0392x_{29} - 0.0367x_{32} - \\
 & 0.8025x_{48} - 0.8744x_{51} - 0.5365x_{56} - 0.3148x_{59} - \\
 & 1.1682x_{77}
 \end{aligned} \tag{5-3}$$

对于 58 肽组，似回归模型 ($PC = 7$) 如下：

$$\begin{aligned}
 A_{58} = & 1.2542 - 0.7320x_1 + 0.0505x_2 + 0.0221x_3 + \\
 & 4.3081x_5 + 1.4426x_6 - 0.7268x_9 - 0.0421x_{14} + \\
 & 0.0788x_{15} + 0.3018x_{17} - 0.0716x_{18} + 0.1915x_{21} + \\
 & 0.4437x_{22} + 0.2768x_{26} - 0.0960x_{28} + 0.2372x_{29} + \\
 & 0.0343x_{32} + 0.0225x_{47} + 0.7201x_{48} + 0.1814x_{51} + \\
 & 1.9970x_{56} - 0.5476x_{59} - 1.0070x_{77} - 0.2874x_{59} + \\
 & 0.3275x_{77} - 1.3599x_{59}
 \end{aligned} \tag{5-4}$$

5.1.5 结果比较

表 5-7 列出 Zaliani 报告^[3]中由 t-scores, z-scores, MS-WHIM 等方法给出的生物活性估计结果以及本文 MHDV 方法结果。显然 本文方法无论是 PCR 或是 PLSR 的结果都优于文献中给定的三种方法。为了结果比较的公正性，表中 r 和 $RMSE$ 是利用文献中给定模型估计的生物活性值与原实验生物活性统一按照一元线性回归方法重新作了统计分析，以使各方法均在同一水平上进行比较。应当指出，这些文献方法都是首先利用 20 种天然氨基酸的各种性质（包括多种物理化学性质和与 3D 结构相关的性质）进行主成分分析，然后对这些主成分进行一系列比较复杂的推导而获

得的一组称之为 z -scores^[6, 9], t -scores^[8] 和 MS-WHIM^[7, 10] 等的所谓 3D 结构分子描述子。然而, 本文提出的 MHDV 方法计算公式是非常简单且只需要 2D 拓扑结构信息, 不需要任何 3D 结构信息或分子的任何物理化学性质, 更加简便, 且结果更优。

表 5-7 两组二肽 QSAR 模型结果比较

方法	$n = 48$				$n = 58$			
	r	RMSE	q	RMSV	r	RMSE	q	RMSV
本文 PCR 方法	0.9625	0.169	0.9273	0.235	0.9293	0.367	0.8802	0.474
本文 PLS 方法	0.9651	0.170	0.9285	0.233	0.9340	0.355	0.8758	0.483
t -scores 法 ^[8]	0.8442	0.340			0.8395	0.540		
z -scores 法 ^[6, 9]	0.9079	0.261			0.8774	0.476		
MS-WHIM 法 ^[7, 10]	0.9079	0.260			0.8643	0.500		

5.2 MHDV 构建取代苯生物降解模型

目前, 已知有机物种类约 700 多万种, 并仍以每年数千种的速率上升, 这些有机物已经并正在通过各种途径进入环境, 现已发现的就有 10 万多种。从环境学的角度来看, 需要考虑这些有机化合物的生物毒性和可降解性两大特性。对于那些已进入环境中的有机物和那些将投入生产和使用的有机物, 其生物毒性和可降解性至关重要。如果通过实验的方法对每种有机物进行危险性和可降解性评价是一个耗时费钱的过程, 是不可取的^[11]。然而, 至今为止, 对于有机物分子结构与可降解性的定量关系研究报道很少^[12~15], 且大多数为简单的单变量模型。王世明等运用神经网络方法和简单化学基团描述子研究了一类芳香化合物的生物可降解性^[11]。本节作者将应用 MHDV 分子描述子表征 51 个取代苯化合物的分子结构, 进而以 PCR 方法建立其生物可降解性定量

相关模型。

5.2.1 取代苯分子结构表征

从 Pitter 及其合作者报告^[16]中选择 51 个单取代苯和二取代苯化合物及它们的有机底物最大比去除速率 Q_{MAX} 的实验测定结果作为研究对象。这个底物最大比去除速率 Q_{MAX} 的大小是一种生物可降解性的量度, Q_{MAX} 越大, 就表示该底物越容易降解。

应用 MHDV 矢量分子描述子表征这些有机底物分子的分子结构。本研究选择的 51 个生物可降解性底物分子都是一取代与二取代苯化合物, 由 C, H, O, N 和 S 等元素构成, 存在“C1”, “CC”, “N1”, “NN”, “O1”, “OO”, “SS”, “L1”, “C2”和“O2”等 10 种原子属性。在它们的 MHDV 矢量中共有 29 个描述子元素具有非零值 其他 62 个 MHDV 描述子对于所有 51 个化合物样本取值都为零。由 29 个非零描述子构成的 MHDV 矢量能很好地表征 51 个取代苯化合物。在由 29 个非零 MHDV 描述子张成的 29 维空间中, 所有这些样本点在该模式空间中是相互分开的, 所有样本矢量对之间的最小距离为 0.3455, 最大距离为 10.1392。说明 MHDV 对这 51 个化合物的分辨率为 100%。

在这 29 个非零 MHDV 描述子中, 有 12 个描述子分别只有 1 (x_6, x_{22}, x_{29}), 2 (x_{24}) 和 3 ($x_3, x_4, x_{20}, x_{23}, x_{28}$) 个非零样本值。这些描述子的非零样本数相对于总样本数而言较少, 其统计显著性较差, 可以先从数据集中删除这 9 个描述子。因此, 实际上只有 20 个非零 MHDV 描述子进入 PCR 分析。

5.2.2 最佳主成分数的确定

在建立最后的主成分回归模型前, 应当选择最佳主成分数。由于数据集中 51 个有机底物分子的生物可降解性 Q_{MAX} 的实验值范围较宽, 其分布也不均匀。若按从小到大顺序, 以 Q_{MAX} 对化合

物指数 (index) 作图如图 5-6A 所示。由图可知, Q_{MAX} 的分布类似于对数变化曲线, 因此, 可考虑对 Q_{MAX} 进行对数变换以降低测定值范围, 使之分布均匀。由于有多个取代苯化合物的 Q_{MAX} 值等于零, 不能直接进行对数变化, 因此, 考虑先在 Q_{MAX} 上加上数值 3 后, 再作自然对数变换, 使所有化合物的生物降解值经变换后均大于零。变换公式如下:

$$RBD = \ln(Q_{\text{max}} + 3) \quad (5-5)$$

式中 RBD (relative biodegradability) 表示经对数处理后的生物降解性, 称之为相对生物可降解性^[17]。变换后的 RBD 的分布 (参见图 5-6B) 明显比 Q_{MAX} 要均匀得多。

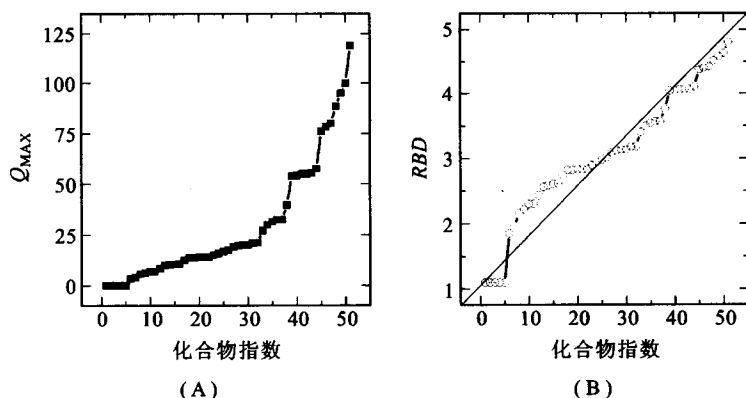


图 5-6 51 个取代苯最大去除速率 Q_{MAX} (A) 与相对生物降解率 RBD (B) 的分布

为了说明生物活性数据预处理的作用与优势, 下面分别以 Q_{MAX} 和 RBD 为因变量, 20 个 MHDV 分子描述子为自变量, 应用主成分回归方法考察主成分对 PCR 模型质量的影响。分别令 $PC = 1, 2, 3, 4, \dots, 20$, 应用主成分分析方法分析不同主成分下, Q_{MAX}

和 RBD 各自与 MHDV 描述子之间的主成分模型及其 LOO 检验的有关统计量比如估计均方根误差 ($RMSE$) 与估计相关系数 (r) 以及 LOO 检验的均方根误差 ($RMSV$) 与检验相关系数 (q) 的变化。均方根误差随主成分数的变化曲线如图 5-7。由图可知, 应用原始底物最大比去除速率 Q_{MAX} 与经对数变换处理的相对生物可降解性 RBD 为因变量所建立的模型的最佳主成分数都是 $PC = 7$ 。然而, 从其模型的相关系数和 LOO 检验相关系数看, 经过对数变换的相对生物可降解性 RBD 与 MHDV 描述子变量之间的相关性更加显著。

两个模型的有关统计量结果如下:

以原始最大比去除速率 (Q_{MAX}) 为因变量的 PCR 回归模型的有关统计量为:

$$PC = 7, n = 51, r = 0.8602, RMSE = 15.05,$$

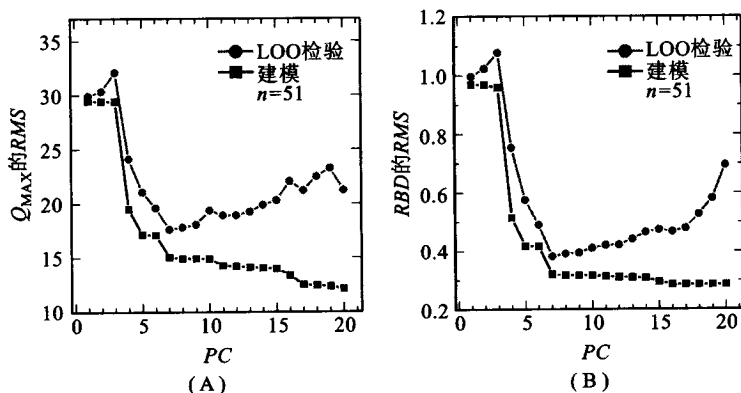


图 5-7. 生物降解性预处理前 (A) 后 (B) PCR 模型中均方根误差随主成分的变化规律

$$q = 0.8047, RMSV = 17.61$$

以生物相对可降解性 (RBD) 为因变量的 PCR 模型的有关统计量为：

$$PC = 7, n = 51, r = 0.9451, RMSE = 0.321, \\ q = 0.9215, RMSV = 0.382$$

显然，经过对数变换后的 PCR 模型更优。这说明，在进行 QSAR 分析时，除了对描述子变量进行优化选择之外，对物质的生物活性 / 理化性质的分布也应该充分考虑，应在所研究的数值范围内保证有充分的均匀分布，否则会影响模型的统计质量。

5.2.3 似回归模型与检验

由图 5-7 可知，最佳主成分数 $PC = 7$ 由此可以求得最佳主成分模型对应的似回归模型。

对于原始最大底物比去除速率 Q_{MAX} 似回归模型 ($PC = 7$) 如下：

$$\begin{aligned} Q_{MAX} = & 31.5919 + 1.4654x_1 + 0.6268x_2 - 16.8623x_5 + \\ & 4.4720x_7 - 0.1275x_8 - 14.5649x_9 - \\ & 12.7930x_{10} + 5.0369x_{11} - 10.4861x_{12} - \\ & 18.5744x_{13} + 2.0764x_{14} - 9.2373x_{15} - \\ & 8.0762x_{16} + 3.0060x_{17} - 7.7344x_{18} - \\ & 12.9834x_{19} - 115.8082x_{21} - 3.2953x_{25} - \\ & 11.2581x_{26} - 2.4725x_{27} \end{aligned} \quad (5-6)$$

$$PC = 7, n = 51, r = 0.8602, RMSE = 15.05, \\ q = 0.8047, RMSV = 17.61$$

对于经对数变换后的相对生物可降解性 RBD ，似回归模型 ($PC = 7$) 如下：

$$\begin{aligned}
RBD = & 3.2482 + 0.2085x_1 + 0.1194x_2 + 0.5781x_3 + \\
& 0.1264x_7 + 0.0054x_8 - 0.5244x_9 - \\
& 0.4829x_{10} + 0.1321x_{11} - 0.3903x_{12} - \\
& 0.5016x_{13} + 0.0713x_{14} - 0.3327x_{15} - \\
& 0.3027x_{168} + 0.0987x_{17} - 0.2860x_{18} - \\
& 0.3544x_{19} - 4.6307x_{21} - 0.1237x_{25} - \\
& 0.4452x_{26} - 0.0917x_{27}
\end{aligned} \tag{5-7}$$

$$\begin{aligned}
PC = 7, n = 51, r = 0.9451, RMSE = 0.321, \\
q = 0.9215, RMSV = 0.382
\end{aligned}$$

由这两个模型估计的最大底物比去除速率 Q_{MAX} 及相对生物可降解性 RBD 与实验值之间的整体相关性绘制在图 5-8 中。显然，未经对数变换的结果比经对数变换的结果差。这也说明能否建立良好的定量结构相关模型，训练集的选择应充分考虑生物活性（因变量）的分布，其分布应该均匀且有较大的极差。如果达不到要求，则应考虑对因变量进行适当的变换。

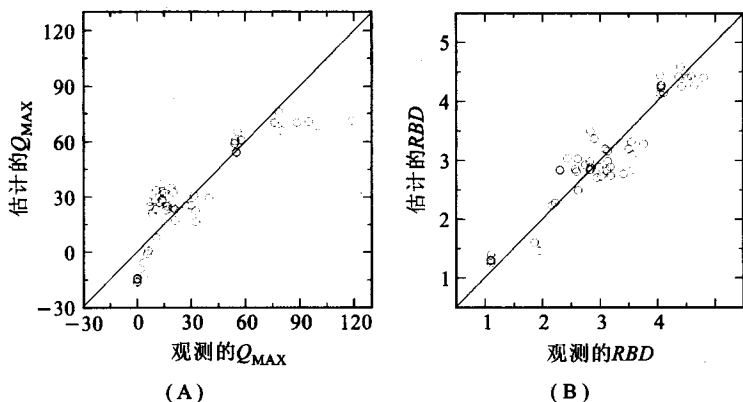


图 5-8 对数变换前 (A) 后 (B) 模型估计结果与实验结果相关图

51 个取代苯化合物、最大底物比消除速率实验值 (Q_{MAX})、相对生物可降解性实验值 (RBD) 与由 (5-7) 式估计的值 (RBD_{M0}) 一并列入表 5-8。

以上说明了数据预处理有利于模型的优化。然而，一个模型的好坏不仅要考察模型的估计能力，而且要考察模型对外部样本的预测能力。模型估计相关性检验与 LOO 检验只能说明模型估计能力与模型稳定性，但如何证明模型的预测能力至今没有统一的标准，但至少应该有外部检验样本的预测结果。然而，如何选择检验样本也没有统一的标准，有人采用随机分组或定制分组的方法来选择训练集和检验集，但为了消除偶然相关，必须选择多个检验集或多次选择检验集。所谓偶然相关就是指某一种分组方法其结果可能很好，而另一种分组对同一个数据集结果相差很大。由于随机分组不易重复，下面采用两种定制分组方案进行试验以讨论偶然相关问题。

第一种定制分组方案从整体数据集 (51 样本) 中选择开始 39 个样本 (1~39 号) 作为内部训练集 余下 12 个样本 (40~51 号) 作为外部检验集 以 39 个训练集数据进行 PCR 分析建立 PCR 潜在 QSAR 模型，并以该 QSAR 模型预测选定的外部检验集样本的生物活性 (RBD)。对 39 个训练集样本进行 PCR 分析 (均方根误差随 PC 的变化见图 5-9A) 可知 最佳主成分数 PC 与全部 51 个样本 PCR 分析的主成分一致，即 $PC=7$ 由此构建的 PCR 模型统计参数及相应的似回归模型 ($PC=7$) 如下：

$$\begin{aligned}
 RBD = & 3.2084 + 0.2439x_1 + 0.1338x_2 + 0.8572x_5 + \\
 & 0.1497x_7 - 0.0180x_8 - 0.4862x_9 - \\
 & 0.5154x_{10} + 0.1180x_{11} - 0.3963x_{12} - \\
 & 0.5666x_{13} + 0.0814x_{14} - 0.3203x_{15} - \\
 & 0.2984x_{16} + 0.0894x_{17} - 0.2908x_{18} - \\
 & 0.3641x_{19} - 4.5134x_{21} - 0.1259x_{25} - \\
 & 0.3469x_{26} - 0.0929x_{27}
 \end{aligned} \tag{5-8}$$

表 5-8 51 个取代苯、 Q_{MAX} 与 RBD 实验值以及各模型计算值

No	化合物	Q_{MAX}	RBD_{OBD}	RBD_{MO}	RBD_{M1}	RBD_{M2}
1	aniline	19.00	3.0910	3.1920	3.2688	3.2373
2	benzaldehyde	119.00	4.8040	4.3957	4.3924	4.4917
3	benzene sulfonic acid	10.60	2.6101	3.0270	3.0954	3.1493*
4	benzoic acid	88.50	4.5163	4.3873	4.3965	4.4301
5	phenol	80.00	4.4188	4.2531	4.2592	4.3515
6	nitrobenzene	14.00	2.8332	2.8429	2.8598	2.9562
7	2-aminobenzoic acid	27.10	3.4045	2.7699	2.8006	2.6093*
8	2-aminophenol	21.10	3.1822	2.7391	2.7536	2.6505
9	2-aminotoluene	15.10	2.8959	3.3700	3.4306	3.2944
10	2-chloroaniline	10.70	2.6174	2.4904	2.4694	2.6081
11	2-chlorophenol	14.00	2.8332	3.4884	3.4052	3.6508*
12	2-hydroxytoluene	54.00	4.0431	4.4384	4.4708	4.3715
13	2-nitroaniline	0.00	1.0986	1.2491	1.2886	1.1546
14	2-nitrobenzaldehyde	13.80	2.8214	2.9802	2.9193	3.0360

续表

No	化合物	Q_{MAX}	RBD_{OBD}	RBD_{M0}	RBD_{M1}	RBD_{M2}
15	2 - nitrobenzoic acid	20.00	3.1355	2.9864	2.9764	2.9879*
16	2 - nitrophenol	14.00	2.8332	2.8767	2.8510	2.9362
17	2 - nitrotoluene	32.50	3.5695	3.0975	3.1596	3.0362
18	phthalic acid	78.40	4.3994	4.5825	4.5621	4.5091
19	pyrocatechol	55.50	4.0690	4.2629	4.2131	4.3080
20	salicylic acid	95.00	4.5850	4.4219	4.3851	4.4091*
21	3 - aminobenzoic acid	7.00	2.3026	2.8308	2.7841	2.6987
22	3 - aminophenol	10.30	2.5878	2.8086	2.7512	2.7494
23	3 - aminotoluene	30.00	3.4965	3.1929	3.1919	3.1024
24	1,3 - benzene disulfonic acid	3.40	1.8563	1.5999	1.6181	1.6853*
25	3 - chloroaniline	6.20	2.2192	2.2827	2.1903	2.3959
26	3 - chlorophenol	31.60	3.5439	3.3156	3.1563	3.4781
27	1,3 - dinitrobenzene	0.00	1.0986	1.3061	1.2186	1.3751
28	3 - hydroxytoluene	55.00	4.0604	4.2658	4.2172	4.2042*

续表

No	化合物	Q_{MAX}	RBD_{OBD}	RBD_{M0}	RBD_{M1}	RBD_{M2}
29	isophthalic acid	76.00	4.3694	4.4200	4.3140	4.3442
30	3 - nitroaniline	0.00	1.0986	1.2984	1.2598	1.2348
31	3 - nitrobenzaldehyde	10.00	2.5649	2.8530	2.7235	2.9046
32	3 - nitrobenzoic acid	7.00	2.3026	2.8400	2.7431	2.8373*
33	3 - nitrophenol	17.50	3.0204	2.7295	2.6221	2.7841
34	3 - nitrotoluene	21.00	3.1781	2.8929	2.8674	2.8377
35	resocinol	57.50	4.1026	4.1382	4.0127	4.1785
36	4 - aminobenzoic acid	12.50	2.7408	2.9272	2.8636	2.8141
37	4 - aminophenol	16.70	2.9806	2.8884	2.8164	2.8454*
38	4 - aminotoluene	20.00	3.1355	3.1574	3.1461	3.0632
39	4 - chloroaniline	5.70	2.1633	2.2358	2.1296	2.3492
40	4 - chlorophenol	39.80	3.7565	3.2810	3.1062*	3.4454
41	1,3 - dinitrobenzene	0.00	1.0986	1.2852	1.1765*	1.3544*
42	hydroquinone	54.20	4.0466	4.1238	3.9824*	4.1640

续表

No	化合物	Q_{MAX}	RBD_{ORD}	RBD_{M0}	RBD_{M1}	RBD_{M2}
43	4 - hydroxybenzoic acid	100.00	4.6347	4.2584	4.1240*	4.2419
44	4 - hydroxytoluene	55.00	4.0604	4.2284	4.1610*	4.1720
45	4 - nitroaniline	0.00	1.0986	1.3912	1.3354*	1.3475*
46	4 - nitrobenzaldehyde	13.80	2.8214	2.8414	2.6987*	2.8933
47	4 - nitrobenzoic acid	19.70	3.1224	2.8250	2.7076*	2.8228
48	4 - nitrophenol	16.00	2.9444	2.7104	2.5847*	2.7650
49	4 - nitrotoluene	32.50	3.5695	2.8425	2.7947*	2.7942*
50	sulfonic acid	4.00	1.9459	1.4610	1.4569*	1.4093
51	4 - toluene sulfonic acid	8.40	2.4336	3.0330	3.0406*	2.9860

注:数据后标有“*”号的表示模型对外部检验样本的预测值,数据来源详见下文。

$$PC = 7, n = 39, r = 0.9507, RMSE = 0.292, \\ q = 0.9112, RMSV = 0.395$$

用上述模型去预测检验集中 12 个样本的 RBD 值, 其预测相关系数 $r_p = 0.9357$, 预测的均方根误差 $RMSP = 0.436$ 。这说明以 39 个训练集构建的 QSAR 模型对所选择的外部检验样本具有良好的预测能力。

第二种定制分组方案从整体数据集 (51 样本) 中大约每 4~5 个样本中选择 1 个样本作为检验集样本, 另 3 个样本作为训练集样本。为了比较选择的训练集样本数仍为 39 个, 检验集样本数仍为 12 个。同样进行主成分分析, 得到主成分模型的均方根误差随主成分的变化曲线 (参见图 5-9B)。由图可知, 最佳主成分数仍是 $PC = 7$ 则由 39 个训练集构建的主成分模型统计量及相应的似回归模型如下:

$$\begin{aligned} RBD = & 3.3941 + 0.1438x_1 + 0.0778x_2 + 0.2824x_5 + \\ & 0.1276x_7 - 0.0093x_8 - 0.5640x_9 - \\ & 0.4791x_{10} + 0.1358x_{11} - 0.3792x_{12} - \\ & 0.3941x_{13} + 0.0520x_{14} - 0.3588x_{15} - \\ & 0.3018x_{16} + 0.0867x_{17} - 0.2825x_{18} - \\ & 0.3060x_{19} - 5.3309x_{21} - 0.1231x_{25} - \\ & 0.4389x_{26} - 0.0916x_{27} \end{aligned} \quad (5-9)$$

$$PC = 7, n = 39, r = 0.9578, RMSE = 0.273, \\ q = 0.8929, RMSV = 0.437$$

用上述模型去预测检验集中 12 个样本的 RBD 值, 其预测相关系数 $r_p = 0.8885$, 预测的均方根误差 $RMSP = 0.447$ 。这说明以 39 个训练集构建的 QSAR 模型对所选择的外部检验样本也具有有良好的预测能力。

为了从整体上考察训练集模型 (5-8) 和 (5-9) 式对训练集样本的估计能力和对外部检验样本的预测能力, 分别应用 (5-8) 和

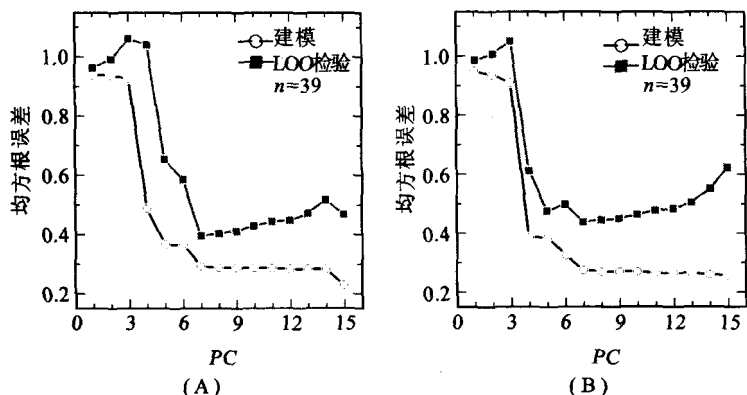


图 5-9 前 39 个样本 (A) 及四抽一样本 (B)

训练集模型 RMS 随 PC 的变化曲线

(5-9) 式计算所有样本的 RBD 值。由(5-8)式估计与预测的 51 个样本 RBD 值(RBD_{M1})以及由(5-9)式估计与预测的 51 个样本 RBD 值(RBD_{M2})一并列入前文给出的表 5-8 中。以这些计算值对原观测实验值作相关图, 得图 5-10。从图可知 39 个训练集所建模型具有高的估计与预测能力。

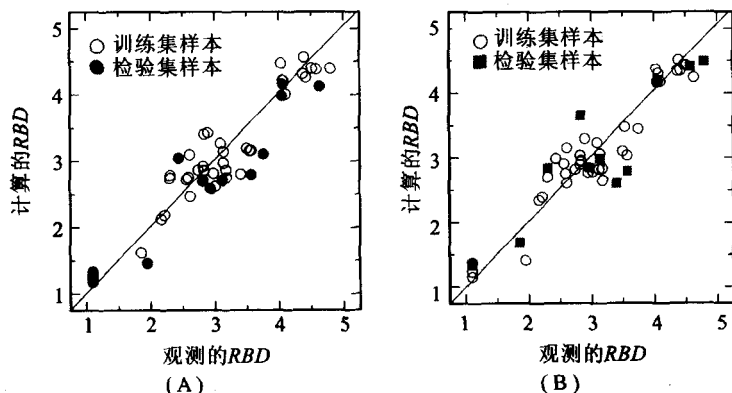


图 5-10 前 39 个样本 (A) 与四抽一样本 (B)

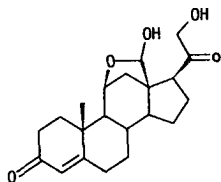
(B) 训练集模型计算 RBD 与实验值相关图

5.3 MEDV-13 研究 31 个甾族化合物的生物活性

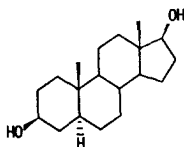
MEDV-13 是另一种基于 13 种原子类型的 MEDV 矢量。与 MHDV 不同, 它使用修正的原子类型电拓扑状态指数即 E-状态指数代替相对电性, 以拓扑距离代替相对距离的分子电性距离矢量^[3,5,18]。在药物分子结构表征及其定量结构-活性相关研究中遇到的最大困难是不像研究结构-性质相关时那样有那么多样本, 即在 QSAR 研究中的样本数一般较少。为了检验 MEDV-13 矢量描述子表征药物分子结构进而进行 QSAR 研究的有效性, 本节首先以一组所谓的“标准”检验物质 (benchmark) 即 31 个甾族化合物为例研究它们的 MEDV-13 矢量分子描述子与相应生物活性之间的定量关系, 并与有关文献进行比较。然后, 再将 MEDV-13 描述子用于一组肠促胰酶肽-A 受体拮抗剂药物分子的 QSAR 研究。在下一章还将专门讨论 MEDV-13 描述子在一系列环加氧酶-2 (简称 COX-2) 选择性抑制剂的生物活性的 QSAR 问题。

5.3.1 31 个甾族化合物的结构表征

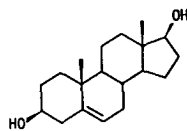
31 个甾族化合物已逐渐成为 QSAR 研究中一组不成文的“标准物质”(benchmark), 它们的分子结构如下:



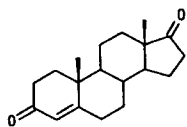
01



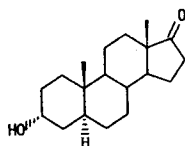
02



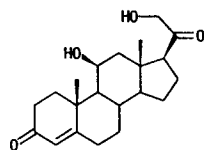
03



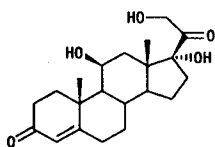
04



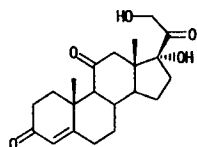
05



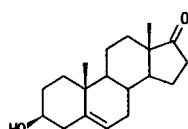
06



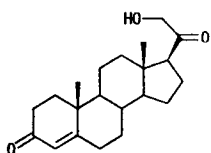
07



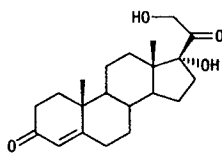
08



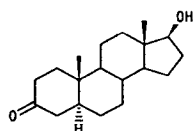
09



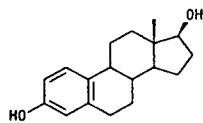
10



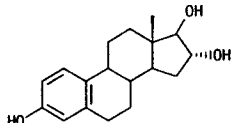
11



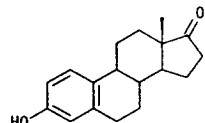
12



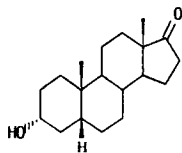
13



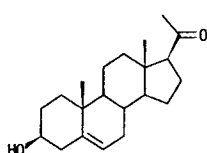
14



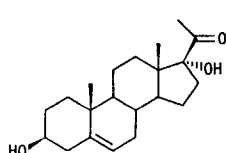
15



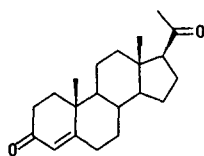
16



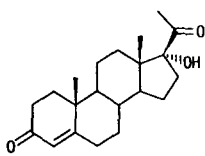
17



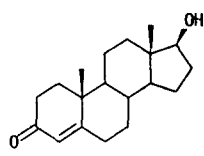
18



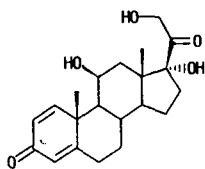
19



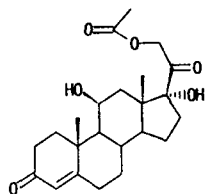
20



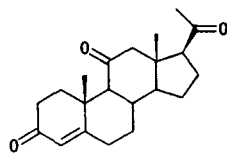
21



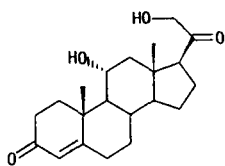
22



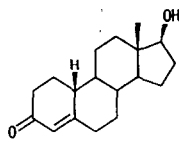
23



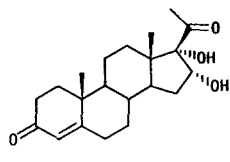
24



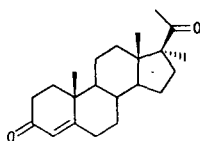
25



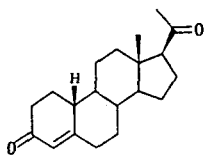
26



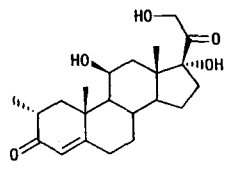
27



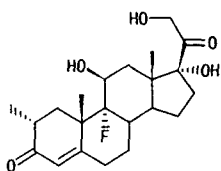
28



29



30



31

对这组化合物的 QSAR, 前人做过相当广泛的研究。Cramer^[19] 以 CoMFA 方法, Good^[20] 利用分子相似性分析, Jain 使用 COMPASS^[21], Bravi^[10] 借 MS-WHIM 描述子以及 Robinson^[22] 通过 SOMFA 等多种不同 QSAR 方法研究过这些标准甾族化合物与一种转移蛋白质即键合皮质醇的球蛋白 (corticosteroid-binding globulin 简称 CBG) 的亲能和预测问题。

这里, 甾族化合物的分子结构用 MEDV-13 描述子表征。由于 31 个分子中没有氮原子, 不存在第 5, 6, 7 和第 8 原子类型以及分子中有氧原子无硫原子, 缺少第 11 和 12 原子类型 因此 在这 31 个标准化合物中实际只有 25 个非零 MEDV-13 描述子 ($x_1, x_2, x_3, x_4, x_9, x_{10}, x_{13}, x_{14}, x_{15}, x_{16}, x_{21}, x_{22}, x_{25}, x_{26}, x_{27}, x_{32}, x_{33}, x_{36}, x_{37}, x_{42}, x_{43}, x_{46}, x_{77}, x_{78}$ 和 x_{81}) 其他所有 66 个 MEDV 描述子取值都为零。按照文献^[12] 定义的原子属性并按公式 (2-8) 式可以计算出这 25 个非零 MEDV-13 描述子。

结果表明 25 个非零 MEDV-13 描述子中 有 5 个描述子 ($x_{13}, x_{25}, x_{36}, x_{46}, x_{81}$) 只有 1 个样本为非零值, 这是因为 31 个化合物中只有 31 号样本含有一个氟原子, 涉及到第 13 原子类型 另有 5 个描述子 ($x_{10}, x_{22}, x_{33}, x_{43}, x_{78}$) 只有 2 个样本为非零值, 这是因为只有 2 个化合物 (第 1 号和第 23 号) 中含有 '—O—' 原子属性 是为第 10 原子类型。这 10 个描述子的非零样本数相对于整个样本数来说均很小, 从统计显著性考虑, 从模型中预先删除。因而, 实际进入 QSAR 分析的只有 15 个 MEDV-13 描述子。

为了说明 MEDV-13 描述子对标准甾族化合物的分辨率, 计算了 MEDV-13 矢量空间 (维数为 $m=15$) 中两两样本间的模式距离。结果表明, 5 号和 16 号以及 6 号和 25 号样本之间的距离为零, 即 $d(5,16)=d(6,25)=0$ 说明 MEDV-13 描述子不能分辨它们。经查这两对化合物只是 H 原子 (No. 5 和 No. 16 样本) 和 OH 基团 (No. 6 和 No. 25 样本) 的空间位置不同。然而, 通过分析它们的生物活性即与

CBG 的亲能值可知 其差值不大 因而不会显著影响 QSAR 结果。

5.3.2 最佳主成分与 PLS 模型

最佳主成分 PC 的选择在基于主成分分析的模型构建方法中具有重要意义。业已证明, 在建模阶段, 估计均方根误差 $RMSE$ 总是随 PC 增加而降低, 相关系数 r 随 PC 增加而升高。然而, 在 LOO 交互检验阶段, 检验 $RMSV$ 和 q 的变化存在一个或多个极值点。一个好的 QSAR 模型必须同时具有良好估计能力和预测能力。然而, 具有良好估计能力的模型不一定具有良好预测能力, 但具有高度预测能力的模型肯定具有好的估计能力, 因此, 最佳主成分的选择主要应由预测阶段决定。虽然, LOO 校验结果只能证明模型的稳定性, 不能完全说明模型对外部样本的预测能力, 但正如文献所证明的, 具有高的 q 值是模型具有良好预测能力的前提。所以, 为了方便, 最佳主成分的选择常以 LOO 检验结果为判据。对于 31 个标准甾族化合物体系, 应用偏最小二乘回归方法进行主成分分析时, 其 $RMSE$ 与 $RMSV$ 随主成分 PC 的变化曲线如图 5-11 所示。由图可知, 最佳主成分数 $PC=6$ 。由 $PC=6$ 构建的 PLS 模型及似回归模型如式 (5-10) 所示。

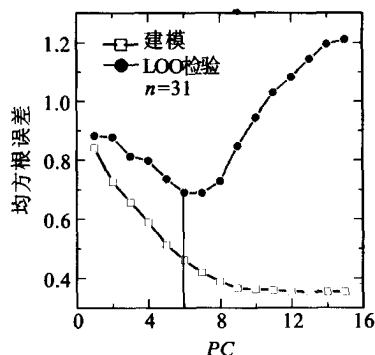


图 5-11 31 个甾族对 CBG 亲和能的 PLSR 模型的 RMS 随 PC 的变化

$$\begin{aligned}
CBG = & -1.3200 + 0.5828x_1 + 0.3596x_2 + 0.1491x_3 + \\
& 0.2384x_4 + 0.0164x_9 - 0.0353x_{14} + \\
& 0.0138x_{15} + 0.0217x_{16} + 0.3999x_{21} + \\
& 0.1053x_{26} + 0.1850x_{27} + 0.0595x_{32} - \\
& 0.3107x_{37} + 0.0199x_{42} + 0.1087x_{77} \quad (5-10)
\end{aligned}$$

$$\begin{aligned}
PC = 6, n = 31, r = 0.9008, RMSE = 0.462, \\
q = 0.7748, RMSV = 0.691
\end{aligned}$$

5.3.3 变量选择与 MLR 模型

通过对最佳主成分下的权矢量和载荷矢量的分析，并建立各个原始描述子变量的投影重要性指标，即 VIP 函数，有可能获得影响 PLSR 模型的主要影响因素，从而对影响生物活性的主要结构因素作出说明和解释。通过 PLSR 分析可知，15 个 MEDV 描述子变量的 VIP 函数大小顺序为： x_{21} (VIP = 1.611)， x_{32} (1.313)， x_1 (1.309)， x_2 (1.298)， x_{14} (1.232)， x_4 (1.164)， x_{15} (1.138)， x_{77} (1.124)， x_{37} (1.093)， x_9 (1.079)， x_{26} (1.011)， x_{42} (0.986)， x_{16} (0.948)， x_{27} (0.942)， x_3 (0.933)。这说明影响对 CBG 亲和能的主要结构因素是 MEDV - 13 描述子 x_{21} ， x_{32} ， x_1 ， x_2 ， x_{14} ， x_4 等对应的有关原子属性与类型。因为 PLSR 模型是关于主成分的隐含模型，而主成分又是原 MEDV 描述子的线性组合，不清楚到底哪些变量是主要影响因素。为了更进一步证实少数主要分子描述子也能得到同 PLSR 最佳主成分模型一样的统计结果，我们进行最佳子集回归变量优化选择与模型分析，也即从 15 个变量中依次选择包含 1, 2, 3, 4, 5, 6, 7 和 8 个描述子变量的子集来构建通常的 MLR 模型。结果表明，各个变量数下最佳量子集模型的均方根误差随变量数的变化曲线如图 5-12 所示，有关统计量结果见表 5-9。由图可知，由描述子 x_1 ， x_2 ， x_4 ， x_{21} 和 x_{32} 等 5 个变量构成的子集模型是最佳模型。

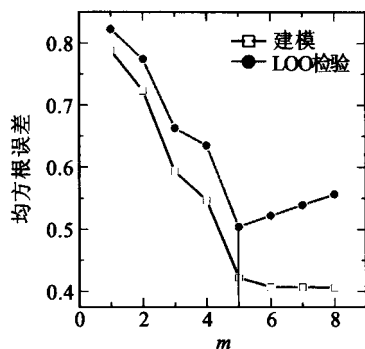


图 5-12 31 个鲎族对 CBG 亲和能的 MLR 模型的 RMS 随变量数的变化

表 5-9 不同变量组合下最佳 ML 模型的有关统计量与 MEDV 描述子

m	r^2	r	RMSE	q^2	q	RMSV	MEDV 描述子
1	0.4525	0.6727	0.788	0.4062	0.6373	0.822	x_{21}
2	0.5392	0.7343	0.723	0.4730	0.6877	0.774	x_2, x_{21}
3	0.6904	0.8309	0.593	0.6146	0.7840	0.663	x_2, x_{16}, x_{21}
4	0.7358	0.8578	0.547	0.6480	0.8050	0.635	$x_2, x_{16}, x_{21}, x_{32}$
5	0.8431	0.9182	0.422	0.7776	0.8818	0.504	$x_1, x_2, x_4, x_{21}, x_{32}$
6	0.8536	0.9239	0.407	0.7651	0.8747	0.522	$x_1, x_2, x_4, x_{21}, x_{32}, x_{37}$
7	0.8536	0.9239	0.407	0.7510	0.8666	0.539	$x_1, x_2, x_4, x_{21}, x_{27}, x_{32}, x_{37}$
8	0.8550	0.9247	0.406	0.7362	0.8580	0.556	$x_1, x_2, x_4, x_{15}, x_{21}, x_{27}, x_{32}, x_{37}$

将 PLSR 最佳主成分下各原始变量的重要性指标 VIP 函数和 MLR 分析中的最佳子集进行比较可以看出，VIP 值排在前 6 位的 6 个重要描述子，除 x_{14} 之外其他 5 个 MEDV-13 描述子全部进入了最佳子集模型。这说明了 PLSR 方法中定义的 VIP 函

数与 MLR 方法中定义的最佳子集变量的一致性。

在 5 变量最佳组合下的回归方程与有关统计量如下：

$$\begin{aligned} \text{CBG} = & -(2.97240 \pm 0.92787) + (0.88298 \pm 0.22983)x_1 + \\ & (0.41455 \pm 0.06958)x_2 + (0.42349 \pm 0.06854)x_4 + \\ & (0.53453 \pm 0.05386)x_{21} + (0.10940 \pm 0.03002)x_{32} \quad (5-11) \\ m = & 6, n = 31, r = 0.9182, \text{RMSE} = 0.422, \\ q = & 0.8818, \text{RMSV} = 0.504 \end{aligned}$$

显然, 5 变量模型估计能力和稳定性都不比 6 个主成分的 PLSR 模型差。由 (5-11) 式估计的 CBG 亲和能数据 (CBG_{M0})、LOO 检验值 (CBG_{LOO}) 与实验观测值 (CBG_{OBS}) 一并列入表 5-10 中。要注意的是, 这里的生物活性 CBG_{OBS} 是以甾族化合物对 CBG 的相对键合亲和能的负对数表示的, 即 $-\lg(\text{RBA})$ 。

表 5-10 31 个甾族化合物及对 CBG 亲和能的
实验值 (CBG_{OBS})、估计值 (CBG_{M0}) 与 LOO 检验值 (CBG_{LOO})

No	化合物	CBG_{OBS}	CBG_{M0}	CBG_{LOO}
1	aldosterone	6.279	6.628	6.686
2	androstanediol	5.000	5.630	5.781
3	androstenediol	5.000	4.732	4.662
4	androstenedione	5.763	6.772	6.953
5	androsterone	5.613	5.652	5.656
6	corticosterone	7.881	7.335	7.288
7	cortisol	7.881	7.473	7.419
8	cortisone	6.892	6.525	6.382
9	dehydroepiandrosterone	5.000	4.781	4.713
10	deoxycorticosterone	7.653	6.759	6.645
11	deoxycortisol	7.881	7.364	7.290

续表

No	化合物	CBG_{OBS}	CBG_{M0}	CBG_{LOO}
12	dihydrotestosterone	5.919	5.968	5.973
13	estradiol	5.000	4.976	4.970
14	estriol	5.000	5.201	5.274
15	estrone	5.000	5.263	5.364
16	etiocholanolone	5.225	5.652	5.697
17	pregnenolone	5.225	5.070	5.035
18	17-hydroxypregnenolone	5.000	5.560	5.668
19	progesterone	7.380	7.114	7.064
20	17-hydroxyprogesterone	7.740	7.516	7.493
21	testosterone	6.724	6.819	6.833
22	prednisolone	7.512	7.890	7.975
23	cortisol 21-acetate	7.553	7.482	7.434
24	4-Pregnene-3,11,20-trione	6.779	7.397	7.472
25	epicorticosterone	7.200	7.335	7.347
26	19-Nortestosterone	6.114	5.791	5.730
27	16a,17-Dihydroxy-4-pregnene-3,20-dione	6.247	6.372	6.404
28	17-Methyl-4-pregnene-3,20-dione	7.120	7.043	6.722
29	19-Norprogesterone	6.817	6.123	5.963
30	11b, 17, 21-Trihydroxy-2a-methyl-4-pregnene-3,20-dione	7.688	8.135	8.214
31	11b, 17, 21-Trihydroxy-2a-methyl-9a-fluoro-4-pregnene-3,20-dione	5.797	5.524	5.422

5.3.4 结果比较分析

前已说明, LOO 检验结果只能说明模型的稳定性, q 的大小也只是模型具有预测能力的必要条件, 为了说明基于 MEDV 描述子的 QSAR 模型对外部样本的预测能力, 也为了与有关文献作比较, 本文还按两种定制分组方案选择检验集来检验模型的预测能力。Cramer 等在研究这 31 个甾族化合物时^[19] 首先利用前面 21 个甾族化合物建立 QSAR 模型并进行交互检验实验。并强调说明只有当这 21 个化合物集的 QSAR 模型确定以后, 才可再用所建模型去预测后面 10 个甾族化合物的生物活性 (CBG 亲和能)。本节采用相同的方法, 选择前面 21 个样本为训练集 后 10 个化合物为检验集, 以训练集建立的 5 变量模型如下:

$$\begin{aligned} \text{CBG} = & -(4.33508 \pm 1.23422) + (0.78296 \pm 0.67041)x_1 + \\ & (0.48137 \pm 0.09917)x_2 + (0.43941 \pm 0.09314)x_4 + \\ & (0.60167 \pm 0.07354)x_{21} + (0.11032 \pm 0.05406)x_{32} \quad (5-12) \\ m = & 5, n = 21, r = 0.9326, \text{RMSE} = 0.414, \\ q = & 0.8719, \text{RMSV} = 0.568 \end{aligned}$$

由 (5-12) 式模型估计的训练集中 21 个样本对 CBG 的亲能的估计值 (CBG_{Ml}) 以及由 (5-12) 式模型预测的外部 10 个检验集样本的预测值 (CBG_{Ml}^*) 列入表 5-11 中 (预测值后标注有“*”号)。应用一元回归方法计算训练集样本估计值与原 21 个样本实验观测值之间的相关系数和均方根误差分别为: $r = 0.9326$ 和 $\text{RMSE} = 0.414$ 这与原文献^[15] 结果为 $r = 0.9471$ 和 $\text{RMSE} = 0.391$ 相比较是基本一致的, 不存在显著性差异。对 21 个训练集样本模型进行 LOO 交互检验得到的 21 个样本的检验预测值与原实验值之间的相关系数和均方根误差分别为: $q = 0.8719$ 和 $\text{RMSV} = 0.568$, 优于原文献结果: $q = 0.8136$ 和 $\text{RMS} = 0.707$ 。这表明本节应用基于分子二维拓扑结构与原子电性得到的 QSAR

模型与原文献应用需要三维活性构象与分子叠合的三维 CoMFA 方法建立的模型具有可比的对内部样本的估计能力或拟合质量，模型稳定性甚至比文献更优。

表 5-11 31 个甾族化合物对 CBG 亲和能实验值、两个训练集模型计算值与 5 个最佳描述子值

No	CBG _{OBS}	CBG _{M1}	CBG _{M2}	x_1	x_2	x_4	x_{21}	x_{32}
1	6.279	6.624	6.742	0.0000	4.4910	-3.2899	16.5144	2.7832
2	5.000	5.563	6.027*	0.3686	9.8221	-3.7198	10.9620	-0.7215
3	5.000	4.623	5.010	0.3518	9.9164	-5.5058	10.6579	-0.7575
4	5.763	6.867	6.592	0.3267	8.2825	-7.3663	15.6036	7.3250
5	5.613	5.597	5.657*	0.3428	8.6430	-6.0023	12.7573	4.2142
6	7.881	7.508	7.400	0.3485	7.6064	-6.6275	17.3463	3.4820
7	7.881	7.771	7.439	0.3450	7.2822	-9.4078	20.1805	2.9187
8	6.892	6.695	6.173*	0.3231	6.0310	-10.7684	19.7789	6.3920
9	5.000	4.688	4.677	0.3269	8.7488	-7.7298	12.4690	4.0758
10	7.653	6.835	6.778	0.3504	9.0557	-5.9688	14.2826	5.1331
11	7.881	7.613	7.277*	0.3469	8.6841	-8.7049	17.8790	5.1167
12	5.919	5.906	6.162	0.3595	8.6706	-4.2018	12.2316	2.4682
13	5.000	4.732	5.110	0.0000	4.2943	-2.1756	12.8824	1.8581
14	5.000	5.023	5.458*	0.0000	3.6353	-2.4907	14.6253	-0.8792
15	5.000	5.056	5.055	0.0000	3.4620	-4.1470	14.6637	6.5670
16	5.225	5.597	5.657	0.3428	8.6430	-6.0023	12.7573	4.2142
17	5.225	4.919	5.424	1.0256	10.9721	-6.4739	10.1981	-1.1058
18	5.000	5.618	5.762*	1.0093	10.5337	-10.3964	14.6547	-1.4247
19	7.380	7.156	7.391	1.0250	10.4405	-6.0984	13.4649	2.1834
20	7.740	7.756	7.639	1.0087	10.0004	-10.0190	17.7649	1.8204

续表

No	CBG _{obs}	CBG _{M1}	CBG _{M2}	x_1	x_2	x_4	x_{21}	x_{32}
21	6.724	6.910	7.022*	0.3516	9.4378	-5.1359	13.9623	2.5628
22	7.512	8.219*	7.708	0.3373	6.6813	-9.9105	21.3338	5.3761
23	7.553	7.869*	7.862	0.5566	7.5711	-10.0895	21.8062	-5.1063
24	6.779	7.496*	7.412*	0.9634	7.3901	-8.2011	17.7418	4.0670
25	7.200	7.508*	7.400	0.3485	7.6064	-6.6275	17.3463	3.4820
26	6.114	5.629*	5.908	0.0000	4.3222	-2.0131	14.0158	3.0355
27	6.247	6.547*	6.673*	1.0031	8.8908	-10.6614	18.1206	-3.6344
28	7.120	6.908*	7.215	3.4995	13.4937	-14.5044	13.4703	2.5057
29	6.817	5.938*	6.317	0.5394	5.4749	-2.7542	13.5168	2.6539
30	7.688	8.411*	8.280*	1.1923	9.6885	-10.1378	18.9197	1.9913
31	5.797	5.639*	5.573	1.1690	9.1005	-13.4834	17.9188	-1.6154

注：数据后标注“*”的表示是模型对外部样本的预测值。

仿文献一样的方法，应用所建训练集模型来预测后面 10 个检验集甾族样本对 CBG 的亲能，并计算预测结果与原实验观测值之间的相关系数 r_p 和均方根误差 $RMSP$ ，结果 $r_p = 0.8902$ 和 $RMSP = 0.538$ 优于文献（参见表 5-12）。

另一方面，本节还利用另一种定制分组方法，按序号均匀地从 31 个甾族化合物中选择 21 个化合物作为训练集样本，余下的 10 个化合物作为检验集样本，并以训练集样本建立 MLR 模型（为讨论方便，简称三选二训练集模型），模型与有关统计量如下：

$$\begin{aligned}
 \text{CBG} = & -(3.10043 \pm 1.05873) + (1.01139 \pm 0.24729)x_1 + \\
 & (0.45751 \pm 0.08393)x_2 + (0.48762 \pm 0.08022)x_4 + \\
 & (0.55823 \pm 0.05979)x_{21} + (0.06207 \pm 0.03711)x_{32} \quad (5-13)
 \end{aligned}$$

$m = 5, n = 21, r = 0.9354, RMSE = 0.380,$

$$q = 0.8815 \quad RMSV = 0.528$$

这个三选二训练集模型与模型(5-12)式相比没有显著性差异,均方根误差无论建模还是检验都更小。为了说明该模型的预测能力,用其预测 10 个外部检验集样本对 CBG 的亲能和,结果表明,预测值与原实验值之间的相关系数和均方根误差分别为: $r_p = 0.8543$ 和 $RMSP = 0.612$ 。与由(5-12)式的预测结果相比稍差,但仍然优于有关文献结果(参见表 5-12)。由(5-13)模型估计 21 个训练集与预测 10 个检验集对 CBG 亲和能结果 (CBG_{M2})也列入前面的表 5-11 中。

由(5-12)和(5-13)式训练集模型对 21 个训练集与 10 个检验集对 CBG 亲和能的计算结果 (CBG_{M2})对相应实验观测结果的相关情况如图 5-13 所示。

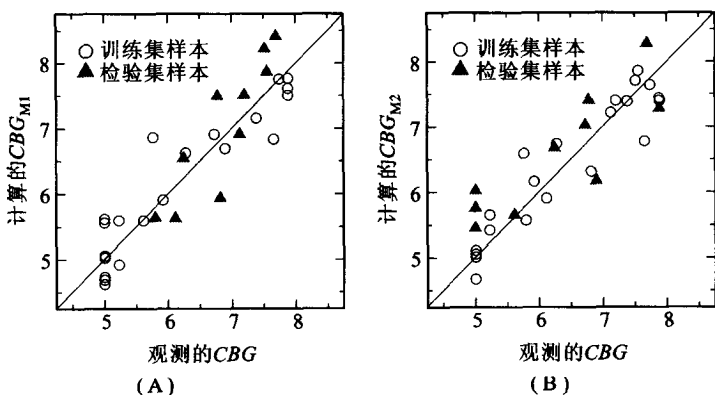


图 5-13 前 21 个(A)与三选二(B)训练集模型计算值与实验 CBG 亲和能相关图

从 PLSR 分析得到的 VIP 函数大小与最佳子集回归(BSR)方法变量选择结果可知,进入最终 QSAR 模型的 5 个 MEDV-13 分

子描述子变量 (x_1, x_2, x_4, x_{21} 和 x_{32}) 是影响甾族化合物对 CBG 亲和能的最主要描述子, 代表着甾族化合物分子对 CBG 相互作用的 5 种主要相互作用, 这些相互作用是在 5 种原子类型 (第 1, 2, 3, 4, 9 原子类型) 之间产生的 (参见 MEDV - 13 矢量中各序号描述子与原子类型的相互关系, 即表 2-7) 即 $x_1 = m_{1-1}$ (第 1 原子类型与第 1 原子类型之间), $x_2 = m_{1-2}$ (第 1 原子类型与第 2 原子类型之间), $x_4 = m_{1-4}$ (第 1 原子类型与第 4 原子类型之间), $x_{21} = m_{2-9}$ (第 2 原子类型与第 9 原子类型之间) 和 $x_{32} = m_{3-9}$ (第 3 原子类型与第 9 原子类型之间)。在 31 个甾族化合物中的这 5 种原子类型分别对应甾族化合物分子中 5 种子结构碎片: CH_3- 、 $>\text{CH}_2$ 、 $>\text{CH}-$ 或 $>\text{C}=\text{}$ 、 $>\text{C}<$ 和 $\text{O}=\text{}$ 或 $\text{HO}-$ 。图 5-14 以第 5 号甾族化合物 androsterone 为例解释了 5 个 MEDV - 13 描述子与分子中存在的各原子类型之间的关系。

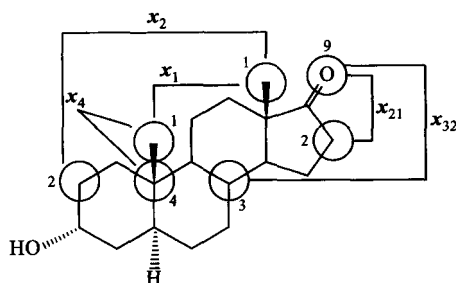


图 5-14 第 5 号化合物 (androsterone) 的结构、原子类型 (黑体数字) 与 MEDV 描述子 (x_i)

为了比较 表 5-12 列出了本文与有关文献的结果。文献结果包括来自 Norinder 等报告^[23] 的 CoMFA 以及 CoMFA (FFD)^[22,23] 来自 Jain 等报告^[21] 的 COMPASS^[21,22] 来自 Bravi 等报告^[10] 的 MS-WHIM^[10,22] 来自 Robinson 等报告^[22] 的相似性分

析以及 SOMFA 等多种 3D QSAR 方法。为了方便, 表 5-12 中所列计算活性数据全部取自 Robinson 等的报告^[22] 并省略了所有负号。表 5-12 结果表明, 预测过程中 MEDV-13 方法的相关系数是所有方法中最高的, 而预测均方根误差则是最低的。这主要是因为 MEDV-13 方法预测 31 号样本时误差是所有方法中最底的, 其他文献方法如两种 CoMFA 方法、COMPASS 方法等对这个含 F 化合物的预测误差都很高。但 MEDV-13 方法对于第 22, 24, 26, 29 和 30 号样本有较大的预测误差, 因此综合起来看 MEDV-13 方法所建模型的质量基本上与当前流行的 3D 方法相当。

表 5-12 各种 QSAR 方法预测 10 个甾族化合物生物活性的结果比较

甾族 化合物	测定 的活性	CoMFA	CoMFA (FFD)	Similarity analysis	COMPASS	MS-WHIM	SOMFA	MEDV
22	7.512	8.084	7.883	7.453	7.062	7.300	7.279	8.219
23	7.553	7.666	7.430	7.022	7.729	8.332	7.034	7.869
24	6.779	6.538	6.642	6.939	6.462	6.821	6.925	7.496
25	7.200	7.804	7.705	7.146	7.466	7.445	7.232	7.508
26	6.144	6.396	6.495	5.908	5.994	6.121	5.744	5.629
27	6.247	7.346	6.962	7.046	6.383	6.901	6.800	6.547
28	7.120	7.010	6.848	6.569	6.625	6.532	6.603	6.908
29	6.817	6.864	6.816	6.850	7.403	6.838	6.692	5.938
30	7.688	7.970	7.767	7.539	7.741	7.860	7.345	8.411
31	5.797	8.005	7.793	7.457	7.779	7.491	7.283	5.639
RMSP		0.837	0.716	0.640	0.705	0.662	0.585	0.538
r_p		0.3659	0.3929	0.3429	0.3924	0.5265	0.4379	0.8902

5.4 MEDV-13 研究 CCK 的生物活性

5.4.1 CCK 化合物的结构表征

肠促胰酶肽 (cholecystokinin 简称 CCK) 是一类肽, 其功能是在周边作为消化荷尔蒙, 在中枢神经系统作为神经调节剂。业已证明, CCK 有两种受体亚型, CCK-A 受体亚型在外周靶器官比如胰脏、胆囊、回肠等占有优势, 但也在离中枢神经系统较远的区域发现过。Tokarski 和 Hopfinger 等^[24] 研究了一系列 CCK-A 受体拮抗剂的三维分子形状描述子与生物活性的定量关系。本节选择 24 个 CCK-A 受体拮抗剂 (以下简称 CCK 分子) 及其生物活性 (IC_{50}) 作为研究对象。这些化合物的母体分子骨架结构如图 5-15 所示。

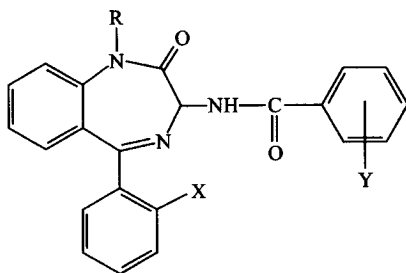


图 5-15 24 个 CCK-A 受体拮抗剂分子的母体结构

这些化合物的分子结构用 MEDV-13 分子描述子表征。由于在所有 CCK 分子中都不存在第 8, 11 和第 12 原子类型原子, 而在这 24 个化合物中真正的非零 MEDV 描述子只有 47 个, 其他所有 44 个 MEDV 描述子对于这 24 个 CCK-A 受体拮抗剂都为零。按照第二章中定义的 43 种原子属性并按公式 (2-8) 式可以

计算出这些化合物所有 47 个非零 MEDV - 13 描述子。计算结果表明 在这 47 个非零 MEDV 描述子中, $x_4, x_5, x_{49}, x_{55}, x_{67}$ 和 x_{85} 等 6 个描述子只有一个非零样本, $x_{10}, x_{17}, x_{22}, x_{28}, x_{33}, x_{40}, x_{48}, x_{51}, x_{60}$ 和 x_{78} 等 10 个描述子只有两个非零样本, $x_{16}, x_{27}, x_{39}, x_{42}$ 和 x_{46} 等 5 个描述子只有三个非零样本, 这 21 个 MEDV 描述子由于非零样本数少, 常常是不显著的, 在进行变量选择或模型建立前可预先删除。因此, 实际进入模型分析阶段的只有 26 个 MEDV 描述子。

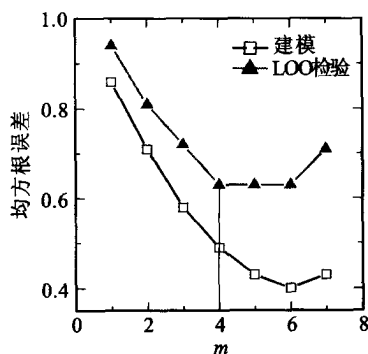
为了说明 MEDV - 13 对 CCK - A 受体拮抗剂分子结构的分辨率, 计算了 MEDV - 13 矢量空间中两两样本 (每个样本由 26 个 MEDV 描述子表示) 间的模式距离。结果表明, 所有样本间的模式距离都大于零, 除 $d(4, 9) = 0.3975$ 外, 其他距离都大于 0.5。这说明 26 个 MEDV 描述子变量能完全分辨 CCK 分子的结构。

5.4.2 变量选择与最佳子集

前面已经表明, 当变量数较大, 而样本数又较小时, 常常采用基于主成分的方法构建关于主成分的定量构效关系模型。当然, 也可以先根据一定规则选择少数优化变量后再进行模型分析。本节将应用最佳子集回归 (BSR) 方法对变量数分别等于 1, 2, 3, 4, 5, 6, 7 时的描述子变量子集进行优化选择, 得到各个变量数下的最佳子集模型, 然后再从这些子集模型中选择全局优化的最佳模型。当变量数分别为 $m = 1, 2, 3, 4, 5, 6, 7$ 时, 经 BSR 选择的最佳变量组合以及有关统计量结果一并列入表 5-13 中。由表 5-13 可知, 在建模阶段, 估计 RMSE 总是随 m 增加而降低, 而 r 却随 m 增加而升高。然而, 在 LOO 交互检验阶段, RMSV 和 q 的变化存在一个或多个极值点。因此模型质量好坏的判别主要由 LOO 交互检验过程决定。对于 CCK 受体拮抗剂体系, 其估计 RMSE 和检验 RMSV 随变量数的变化如图 5-16 所示。由图 5-16 可方便地看出, 变量数 $m = 4$ 时的模型为最优。

表 5-13 CCK 体系各变量数下模型统计量及相应描述子

m	r^2	r	RMSE	q^2	q	RMSV	MEDV 描述子
1	0.4293	0.6552	0.86	0.3277	0.5725	0.94	x_{57}
2	0.6120	0.7823	0.71	0.5015	0.7081	0.81	x_{15} x_{59}
3	0.7443	0.8627	0.58	0.6115	0.7820	0.72	x_{14} x_{15} x_{29}
4	0.8159	0.9033	0.49	0.7036	0.8388	0.63	x_{14} x_{26} x_{29} x_{59}
5	0.8600	0.9273	0.43	0.7167	0.8466	0.63	x_{14} x_{15} x_{26} x_{29} x_{59}
6	0.8740	0.9349	0.40	0.7177	0.8472	0.63	x_{14} x_{15} x_{25} x_{26} x_{29} x_{59}
7	0.8595	0.9271	0.43	0.6624	0.8139	0.71	x_9 x_{14} x_{26} x_{29} x_{36} x_{59} x_{91}


 图 5-16 CCK 体系模型均方根
误差随变量数的变化曲线

由表 5-13 可知 最佳子集中的 4 个 MEDV-13 描述子分别是： x_{14} , x_{26} , x_{29} 和 x_{59} 它们分别表征原子类型 2 与 2, 3 与 3, 3 与 6 以及 6 与 9 之间的相互作用。这也就是说, 对于 CCK 分子 这些原子类型 (参考表 2-6) 相应的子结构碎片 “ $-\text{CH}=\text{}$ ”、“ $>\text{CH}$

“=”, “-N=”或“-NH-”和“O=”是影响 CCK 分子生物活性的最主要因素, 这些子结构碎片都包含在母体结构中, 三个取代基主要是改变母体结构中某些原子的原子类型, 比如 R 取代基位置如果有非氢原子取代, 则母体上的 N 原子类型由 2 (-NH-) 变成了 3 (-N=)。这 4 个重要的 MEDV 描述子与 4 个原子类型的关系示于图 5-17 中。

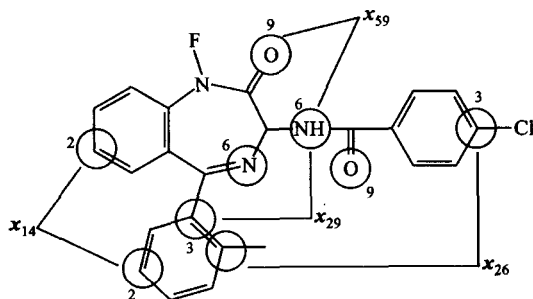


图 5-17 CCK 分子 (No. 62) 中的原子类型和最佳 MEDV 描述子的关系

5.4.3 多元线性回归模型

应用优化选择的 4 个 MEDV-13 描述子变量, 以 MLR 方法建立 CCK 化合物生物活性 (以 pIC_{50} 表示) 的多元线性回归模型, 结果如下:

$$\begin{aligned}
 pIC_{50} = & (59.37589 \pm 7.54301) - (0.31463 \pm 0.05533)x_{14} - \\
 & (2.87001 \pm 0.49601)x_{26} + (5.62691 \pm 0.99582)x_{29} + \\
 & (2.38820 \pm 0.35333)x_{59} \quad (5-14) \\
 m = 4, n = 24, r = 0.9033, RMSE = 0.49, \\
 q = 0.8388, RMSV = 0.63
 \end{aligned}$$

由上式估计的 24 个 CCK 化合物的拮抗活性 (pIC_{EST}) 与实验

观测活性 (pIC_{50}) 一起列入表 5-14 中。计算值与实验值之间的整体相关情况见图 5-18。

为了说明模型对外部样本的预测能力，仅有 LOO 检验是不够的。前面章节中讨论了两种定制分组的方法来说明模型的预测能力，这里再说明一种也能说明预测能力的所谓 LNO 交互检验方法。所谓 LNO 就是 leave N-sample out 的简称，即在进行交互检验时，先将数据集分成若干组，每组包括 N 个样本，然后抽出第一组作为检验集，余下的作为训练集，以训练集建立多元线性回归模型，并以这个模型去预测抽出的那 N 个检验样本，然后将检验集放回数据集，抽出第二组样本，余下的作为训练集建立模型，并预测被抽出的检验样本；依此类推，直到所有组都被抽出一一次为止。在 CCK 体系中，先将 24 个样本分成 6 组，每组 4 个样本，按上述方法进行 LNO 检验。6 次建模与预测的相关统计量结果参见表 5-15。LNO 方法预测的 pIC_{50} 值列入也列入前面的表 5-14，预测值对实验值的相关情况如图 5-19。

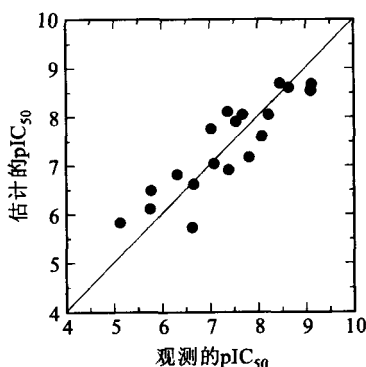


图 5-18 CCK 体系 4 变量模型估计
与实验活性相关图

表 5-14 24 个 CCK 受体拮抗剂化合物及其实验活性(pIC_{50})、估计活性(pIC_{EST})、LNO 预测活性(pIC_{PRD})与 4 个最优 MEDV-13 分子描述子

No	R	X	Y	pIC_{50}	pIC_{EST}	pIC_{PRD}	x_{14}	x_{26}	x_{29}	x_{59}
12	F	H	<i>p</i> -Cl	8.22	8.04	7.99	62.1874	4.6713	-0.5403	6.4159
62	F	CH ₃	<i>p</i> -Cl	8.64	8.61	8.54	55.5718	6.2031	0.0099	6.5026
64	H	CH ₃	<i>p</i> -Cl	8.08	7.60	7.42	56.1486	6.6016	0.9853	8.6693
65	H	H	<i>p</i> -Cl	7.39	6.91	6.76	62.8340	4.9807	0.3635	8.5619
69	H	H	<i>o</i> -Cl	5.12	5.83	6.10	64.9807	5.5858	0.6420	8.6588
70	H	CH ₃	<i>o</i> -Cl	5.77	6.49	6.84	58.3191	7.2255	1.2685	8.7667
72	F	H	<i>p</i> -CF ₃	7.82	7.17	6.99	61.2203	4.5259	-0.8612	6.3255
73	F	H	<i>p</i> -CH ₃	7.68	8.05	7.97	61.9410	4.6519	-0.5659	6.4061
76	H	H	<i>m</i> -Cl	7.09	7.04	7.15	62.5362	5.0532	0.4214	8.5944
77	F	H	<i>p</i> -OCH ₃	7.02	7.76	7.89	62.3920	4.6571	-0.5836	6.4228
81	H	H	3,4-di-Cl	7.54	7.90	7.97	55.6840	5.7809	0.5926	8.6669
87	H	H	<i>p</i> -SCH ₃	6.66	6.62	6.73	64.9769	5.1335	0.5535	8.6656
88	H	H	<i>p</i> -F	6.32	6.82	6.77	60.8345	4.8121	0.1056	8.4576

续表

No	R	X	Y	pIC ₅₀	pIC _{EST}	pIC _{PRD}	x ₁₄	x ₂₆	x ₂₉	x ₅₉
90	F	CH ₃	<i>m</i> -Br	8.46	8.70	8.74	55.8983	6.3977	0.1709	6.5706
92	H	H	<i>p</i> -CF ₃	6.62	5.73	5.58	61.8628	4.8158	-0.0395	8.4330
93	F	CH ₃	<i>p</i> -Br	9.10	8.56	8.60	56.2726	6.2715	0.0868	6.5309
95	F	CH ₃	<i>p</i> -I	9.10	8.54	8.41	56.6007	6.3038	0.1227	6.5440
97	F	CH ₃	<i>p</i> -CN	7.37	8.11	8.44	52.1378	5.8084	-0.5969	6.2089
103	H	H	<i>p</i> -OH	5.75	6.11	6.31	61.3599	4.8555	0.1739	8.7963
104	F	CH ₃	<i>m</i> -I	9.12	8.68	8.53	56.1868	6.4578	0.2227	6.5891
105	H	H	<i>p</i> -CN	6.14	5.78	5.59	59.2104	4.6515	-0.3710	8.1751
108	F	CH ₃	<i>o</i> -I	6.10	6.36	6.13	58.6471	7.5151	0.5236	6.6771
118	F	CH ₃	<i>p</i> -CF ₃	7.70	7.81	7.74	54.6610	6.0018	-0.3243	6.4119
123	H	CH ₃	<i>o</i> -I	5.96	5.54	5.24	59.2313	7.9897	1.5862	8.8763

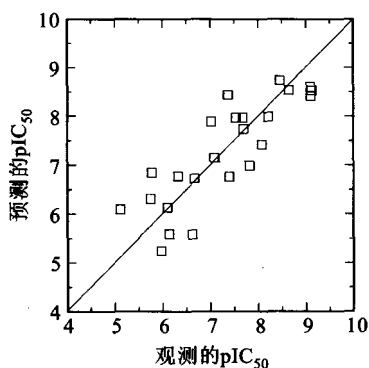


图 5-19 CCK 体系 LNO 交互检验
预测值与实验值相关图

表 5-15 CCK 体系 LNO 交互检验结果

No	训练集	检验集	r	RMSE	q	RMSV	r_p	RMSP
1	5~24	1~4	0.9021	0.50	0.8215	0.68	0.9743	0.47
2	1~4, 9~24	5~8	0.9193	0.42	0.8481	0.58	0.7988	0.85
3	1~8, 13~24	9~12	0.9155	0.50	0.8394	0.69	0.7914	0.49
4	1~12, 17~24	13~16	0.9111	0.46	0.8308	0.64	0.8897	0.63
5	1~16, 21~24	17~20	0.9015	0.45	0.8291	0.59	0.8677	0.75
6	1~20	21~24	0.8939	0.51	0.7542	0.77	0.9598	0.45

参 考 文 献

- [1] Liu S S, Yin C S, Cai S X, Li Z L. A novel MHDV descriptor for dipeptide QSAR studies. J. Chin. Chem. Soc., 2001, 48(2): 253~260
- [2] Liu S S, Yin C S, Wang X D, Wang L S. QSAR studies on dipeptides

based on a combinatorial MHDV - GA - MLR method. J. Chin. Chem. Soc., 2002, 49(6): 1089~1096

- [3] Liu S S, Yin C S, Cai S X, Li Z L. QSAR study of steroid benchmark and dipeptides based on MEDV - 13. J. Chem. Inf. Comput. Sci., 2001, 41 (2): 321~329
- [4] Liu S S, Cui S H, Shi Y Y, Wang L S. Novel variable selection and modeling method based on the prediction for QSAR of cyclooxygenase - 2 inhibition by thiazolone and oxazolone series. Internet Electron. J. Mol. Des., 2002, 1(11): 610~619
- [5] Liu S S, Liu H L, Shi Y Y, Wang L S. QSAR of cyclooxygenase - 2 (COX - 2) inhibition by 2,3 - diarylcyclopentenones based on MEDV - 13. Internet Electron. J. Mol. Des., 2002, 1(6): 310~318
- [6] Collantes E R, Dunn W J III. Amino acids side chain descriptors for quantitative structure - activity relationship studies of peptide analogues. J. Med. Chem., 1995, 38: 2705~2713
- [7] Zaliani A, Gancia E. MS - WHIM scores for amino acids: a new 3D - descriptor for peptide QSAR and QSPR studies. J. Chem. Inf. Comput. Sci., 1999, 39: 525~533
- [8] Cocchi M, Johansson E. Amino acids characterization by GRID and multivariate data analysis. Quant. Struct.-Act. Relat., 1993, 12: 1~8
- [9] Hellberg S, Sjostrom M, Skagerberg B, Wold S. Peptide quantitative structure-activity relationships, a multivariate approach. J. Med. Chem., 1987, 30: 1126~1135
- [10] Bravi G, Gancia E, Mascagni P, Pegna M, Todeschini R, Zaliani A. MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties; a comparative 3D QSAR study in a series of steroids. J. Comput.-Aided Mol. Des., 1997, 11: 79~92
- [11] 王世明, 施汉昌, 钱易. 环境化学, 2000, 19: 48~52
- [12] Boethling R S. Application of molecular topology to quantitative structure - activity relationships. Environ. Toxi. -Col. Chem., 1986, 5: 797~806
- [13] Dearden J C, Nicholson R M. The prediction of biodegradability by the use of quantitative structure-activity relationships: correlation of biological oxygen

- demand with atomic charge difference. *Pestic. Sci.*, 1986, 17: 305~310
- [14] Kompare B. Estimating environmental pollution by xenobiotic chemicals using QSAR (QSBAR) models based on artificial intelligence. *Wat. Sci. Tech.*, 1998, 37: 9~18
- [15] Damborsky J, Schultz W T. Comparison of the QSAR models for toxicity and biodegradability of anilines and phenols. *Chemosphere*, 1997, 34: 429~446
- [16] Pitter P, Chudoba J. Biodegradability of organic substances in the aquatic environment. Boston: CRC press, 1990
- [17] Liu Y, Liu S S, Cui S H, Cai S X. A novel quantitative structure-biodegradability relationship (QSBAR) of substituted benzenes based on MHDV descriptor. *J. Chin. Chem. Soc.*, 2003, 50(2): 319~324
- [18] Liu S S, Yin C S, Cai S X, Li Z L. Molecular structural vector description and retention index of polycyclic aromatic hydrocarbons. *Chemom. Intel. Lab. Syst.*, 2002, 61: 3~15
- [19] Cramer R D III, Patterson D E, Bunce J D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.*, 1988, 110: 5959~5967
- [20] Good A C, So S S, Richards W G. Structure - activity relationships from molecular similarity matrixes. *J. Med. Chem.*, 1993, 36: 433~438
- [21] Jian A N, Koile K, Chapman D. Compass: predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J. Med. Chem.*, 1994, 37: 2315~2327
- [22] Robinson D D, Winn P J, Lyne P D, Richards W G. Self - organizing molecular field analysis: a tool for structure-activity studies. *J. Med. Chem.*, 1999, 42: 573~583
- [23] Norinder U. 3D - QSAR investigation of the Tripos benchmark steroids and some protein tyrosine - kinase inhibitors of styrene type using the TDQ approach. *J. Chemometr.*, 1996, 10: 533-545
- [24] Tokarski J S, Hopfinger A J. Three - dimensional molecular shape analysis - quantitative structure - activity relationship of a series of cholecystokinin-A receptor antagonists. *J. Med. Chem.*, 1994, 37: 3639~3654

第6章

几类 COX-2 抑制剂 QSAR 研究

非甾类抗炎药物通过抑制环氧合酶 (cyclooxygenase) 催化花生四烯酸炎症介质前列腺素, 达到解热、镇痛及治疗多种炎症的目的。环氧合酶有两种结构亚型^[1~3], 即结构型环氧合酶-1(COX-1)和诱导型环氧合酶-2(COX-2)。环氧合酶-1主要存在于正常的组织细胞中, 如分布于血管、胃肠道、肾脏等组织, 催化产生维持正常生理功能所需的前列腺素; 诱导型的环氧合酶-2主要存在于炎症组织细胞中, 在炎症介质的诱导下产生的环氧合酶-2(COX-2)催化花生四烯酸, 产生致炎介质前列腺素。现有的非甾类抗炎药物由于不加选择地抑制两种环氧合酶, 因而在治疗炎症的同时也抑制了正常生理功能所需的前列腺素, 造成胃肠道刺激、损伤和溃疡等毒副作用^[4]。研制高效、低毒、副作用小的非甾类抗炎药物的新途径就是开发选择性的 COX-2 抑制剂。COX-2 的发现 (1990 年) 为非甾类抗炎药物 (nonsteroidal antiinflammatory drugs, NSAIDs) 的研究开辟了一个新方向。1994 年, Vane^[5] 指出 NSAIDs 对炎症的有效治疗作用源于其对 COX-2 的抑制, 而不良反应归于对 COX-1 的抑制。因此选择性抑制 COX-2 就可能有效地治疗炎症, 同时避免或减轻由于抑制维持

人体正常生理功能所必需的 COX - 1 而导致的毒副作用。 1995 年 ,Grossman^[6]利用酶免疫测定法确定了不同 NSAIDs 对两种 COX 的抑制活性和其选择性。研究证明,药物对 COX - 2 抑制的选择性越强,诱发胃溃疡的副作用越弱,且呈良好线性关系;而阿司匹林和吲哚美辛等一些引起较严重胃溃疡的药物,对 COX - 2 的抑制作用较弱,相反对 COX - 1 的抑制作用很强。因此, 1990 年代以来 寻找 COX - 2 选择性抑制剂已成为研究的热点和前沿^[7], 已开发出了大量保留 NSAIDs 治疗炎症药效又避免肠胃副作用的 COX - 2 选择性抑制剂^[8~15]。近几年来,国外一些大的制药公司如美国的 Searle 和 Merck、瑞士的 Ciba - Geigy 和 Roche、日本的 Taisho、Tabacco 和 Fujisawa 以及英国的 Glaxo Wellcome 等都在积极进行开发 COX - 2 选择性抑制剂的研究,且已发现了一些有药用前途的化学实体。目前正在开发研究的 COX - 2 选择性抑制剂从化学结构上可分为二苯基取代芳环类(如 DuP - 697^[16,17]、SC - 58125^[18]、SC - 58635^[19])、甲磺酰苯胺类(如 NS - 398^[20~22]、CGP 28238^[23]等)、二叔丁基取代苯酚类(如 BF389^[24])等多种类型。

定量构效关系(QSAR)技术是对物质生物性质进行分子解释研究中不可缺少的工具^[25]。利用 QSAR 技术,可以识别产生最大化抑制 COX - 2 又同时最小化抑制 COX - 1 作用的药物分子结构所必须具有的分子参数或分子描述子,从而获得更加安全有效的 NSAIDs 的选择性 COX - 2 抑制剂。因而,构建这些化合物的分子结构与生物活性之间的定量相关模型对于研究、设计和开发出高效低毒抗炎药物具有不可估量的重要意义^[26]。然而,尽管有关 COX - 2 抑制剂的研究报道已有近千篇,但 COX - 2 抑制剂抑制生物活性的 QSAR 研究报道较少^[27~31]。基于此,本章应用 MEDV - 13 矢量分子描述子来表征有关 COX - 2 抑制剂的分子结构,并对几类未曾报道过的 COX - 2 抑制剂比如吲哚美辛及其酰胺与酯衍生物、二芳基环戊烯酮类抑制剂等的抑制活性 QSAR

进行研究，获得了一些有意义的结果。

6.1 吲哚美辛及其酰胺与酯

6.1.1 吲哚美辛衍生物的结构表征

选择最近 2000 年 Kalgutkar 等^[32]研究并报道过定量生物活性的吲哚美辛和它的酰胺及酯衍生物 (indomethacin amides and esters 简称 ImAE) 作为研究对象。这些化合物的实验生物活性是两次平行测定的平均值, 以 IC_{50} 表示, 单位是 $\mu M (10^{-6} \text{ mol/L})$ 。本节以 IC_{50} 的负对数 pIC_{50} 作为生物活性指标来研究这些 ImAE 的分子结构与生物活性之间的定量关系。选定的 16 个吲哚美辛衍生物的母体结构如图 6-1, 这些化合物的分子结构以 MEDV-13 表征。

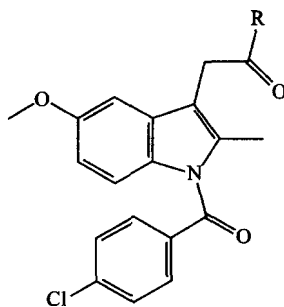


图 6-1 ImAE 的母体结构

对于 ImAE 的 16 个化合物, 由于分子中缺少第 4,8,11 和第 12 原子类型, 因此 91 个 MEDV-13 分子描述子中实际上只有 42 个具有一个或多个非零值。这 42 个描述子能很好地分辨 16 个 COX-2 抑制剂 (ImAE) 的分子结构, 所有 ImAE 化合物对在

MEDV 描述子变量空间中的模式距离都大于零，其中最短距离为 3.062，最大距离达 35.939，说明 MEDV 对这些分子的结构分辨率达 100%，相应距离分布直方图如图 6-2 所示。

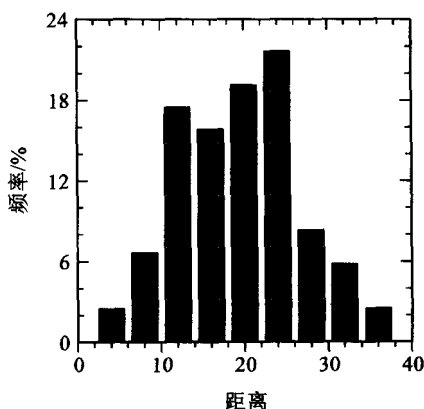


图 6-2 不同距离范围样本分布频率

在这 42 个非零 MEDV 描述子中，描述子 $x_5, x_{17}, x_{28}, x_{49}, x_{51}, x_{52}$ 和 x_{55} 只有 1 个样本具有非零值， x_{56} 和 x_{91} 只有 2 个样本具有非零值，所以这 9 个描述子可以首先从数据集中删去。这样，进入模型分析的实际上只有 33 个描述子。

6.1.2 最佳主成分与重要描述子选择

在 16 个 ImAE 化合物中有 33 个 MEDV-13 分子描述子不能直接应用 MLR 方法建立定量相关模型，可以采用基于主成分分析的 PCR 或 PLSR 方法，也可以先进行变量优化选择再进行模型化。本节一方面应用 PLSR 方法建立关于主成分的 PLSR 模型并进行重要变量影响分析。

应用 PLSR 方法，计算主成分分别等于 $PC = 1, 2, 3, \dots, 14$ 时的 PLSR 模型，应用各主成分模型对 16 个 ImAE 化合物的生物活

性(pIC_{50})进行估计 同时进行 LOO 交互检验 , 计算有关估计与检验统计量如均方根误差 $RMSE$ 和 $RMSV$ 。结果表明 , 不同主成分下 , PLSR 模型的均方根误差的变化曲线如图 6-3 所示。

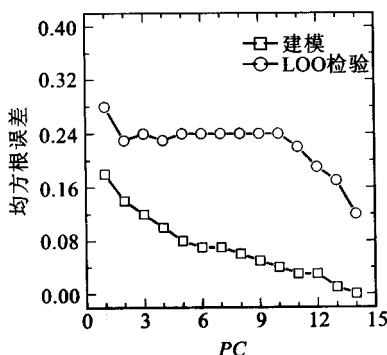


图 6-3 PLS 模型中均方根误差随主成分的变化

由图 6-3 可知 , 最佳主成分数为 $PC = 2$ 。此时 , PLSR 模型的几个重要统计量如下 :

$n = 16, m = 33, PC = 12, r = 0.9514, RMSE = 0.14$ (建模)

$n = 16, m = 33, PC = 12, q = 0.8606, RMSV = 0.23$ (LOO 检验)

由 $PC = 2$ 的 PLSR 模型估计的 ImAE 化合物的 pIC_{50} (记为 pIC_{PLS}) 以及 LOO 交互检验预测的 pIC_{50} (记为 PLS_{LOO}) 与实验 pIC_{50} (记为 pIC_{50}) 值一并列入表 6-1 中 , 估计活性与实验活性总体相关情况如图 6-4 所示。

在 $PC = 2$ 的情况下 , 经 PLSR 分析可知各原始 MEDV-13 描述子变量的重要性大小。在第三章介绍的原始变量的投影重要性指标 variable importance in projection 简称 VIP 是原始变量对生物活性贡献重要性的量度 , VIP 越大 , 重要性就越大。按 (3-29) 式计算得到的 VIP 函数值由大到小的顺序排列为 :

$$x_{19}、x_{14}、x_{21}、x_{66}、x_{22}、x_{77}、x_{29}、x_{67}、x_{18}、x_{15}、x_{78}$$

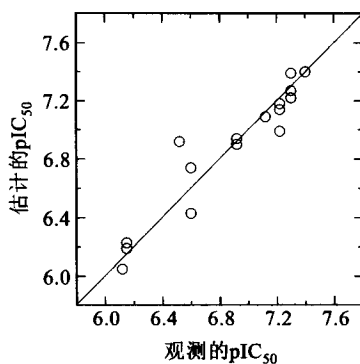


图 6-4 PLS 模型估计与实验 pIC_{50} 相关图

表 6-1 16 个 ImAE 及其实验的、PLSR 模型估计的及 LOO 检验的 pIC_{50} 值

No	—R	pIC_{50}	pIC_{PLS}	PLS_{LOO}
1	—OH	6.12	6.05	6.24
2	—NHCH ₃	6.15	6.23	6.49
3	—OCH ₃	6.60	6.43	6.39
4	—NHCH ₂ CH ₂ OH	6.60	6.74	6.78
5	—NHC ₆ H ₅ (4 - NHCOCH ₃)	6.92	6.94	6.93
6	—OC ₆ H ₅ (4 - OCH ₃)	7.40	7.40	7.21
7	—OC ₆ H ₅ (4 - SCH ₃)	6.52	6.92	7.11
8	—OC ₆ H ₅ (2 - SCH ₃)	7.22	6.99	6.85
9	—OC ₆ H ₅ (4 - F)	7.12	7.09	7.24
10	—O(3 - C ₅ H ₄ N)	7.30	7.39	7.26
11	—NHC ₆ H ₅ (4 - SCH ₃)	6.92	6.90	6.95
12	—NHC ₆ H ₅ (4 - F)	7.22	7.14	7.12

续表

No	—R	pIC ₅₀	pIC _{PLS}	PLS _{LOO}
13	—NH(3—C ₃ H ₄ N)	7.30	7.27	7.14
14	—NH ₂	6.15	6.19	6.29
15	—NHCH ₂ CH ₂ C ₆ H ₅	7.22	7.18	7.13
16	—OCH ₂ CH ₂ C ₆ H ₅	7.30	7.22	7.16

VIP 结果表明, MEDV - 13 描述子 x_{19} 等少数变量对 ImAE 化合物 pIC₅₀ 的影响是最大的。或者说, 33 个非零的 MEDV 描述子中 VIP 函数排前面的少数几个描述子是最重要的。为了更进一步地证明这一点, 对 33 个 MEDV 描述子体系进行 BSR 分析, 结果表明, 不同变量数的各个模型所选择的 MEDV - 13 描述子及有关统计量如表 6-2 所示。

表 6-2 BSR 方法选择的不同变量数最佳子集模型的有关统计量结果

No	r^2	r	RMSE	q^2	q	RMSV	MEDV - 13 描述子
1	0.650 3	0.806 4	0.26	0.554 6	0.744 7	0.30	x_{19}
2	0.770 4	0.877 7	0.21	0.659 3	0.812 0	0.26	$x_{10} x_{19}$
3	0.830 0	0.911 0	0.18	0.758 0	0.870 6	0.22	$x_{14} x_{22} v_{29}$
4	0.886 3	0.941 5	0.15	0.782 4	0.884 5	0.21	$x_{10} x_{18} x_{19} x_{26}$
5	0.932 6	0.965 7	0.11	0.829 1	0.910 6	0.19	$x_1 x_{10} x_{18} x_{19} v_{26}$
6	0.945 0	0.972 1	0.10	0.872 6	0.934 1	0.17	$x_6 x_{10} x_{18} x_{19} x_{26} x_{36}$
7	0.969 5	0.984 6	0.08	0.898 1	0.947 7	0.14	$x_1 x_{10} x_{18} x_{19} x_{26} x_{59} x_{81}$

表 6-2 结果表明, 随着变量数的增加, 估计相关系数 r 与 LOO 检验相关系数 q 都在增大, 估计均方根误差 RMSE 与 LOO 检验均方根误差 RMSV 都在减少, 在变化曲线上不存在明显的转折点或极值点, 似乎不好决定全局最佳变量子集。然而, 根据经验规则, 进入回归模型的变量数一般不能大于样本数的 1/6 或

1/5, 因此最佳变量数只能是 2 或 3。当 $m = 2$ 时, $r = 0.8777$, $RMSE = 0.21$, $q = 0.8120$, $RMSV = 0.26$, 选入的变量是 MEDV - 13 描述子 x_{10} 和 x_{19} ; 当 $m = 3$ 时, $r = 0.9111$, $RMSE = 0.18$, $q = 0.8706$, $RMSV = 0.22$, 选入的变量是 MEDV - 13 描述子 x_{14} , x_{22} 和 x_{29} 。结合 PLSR 结果可知, 相比之下 3 变量模型统计量与 PLSR 结果差异更小, 所选变量也是 PLSR 分析得出的重要影响变量, 因此 3 变量模型为优。这个三变量模型如下:

$$pIC_{50} = (5.382\ 58 \pm 0.219\ 93) + (0.022\ 46 \pm 0.005\ 46)x_{14} + \\ (0.096\ 46 \pm 0.028\ 33)x_{22} + (0.761\ 77 \pm 0.216\ 07)x_{29} \quad (6-1)$$

$m = 3, n = 16, r = 0.911\ 1, RMSE = 0.18$ (建模)

$m = 3, n = 16, q = 0.870\ 6, RMSV = 0.22$ (LOO 检验)

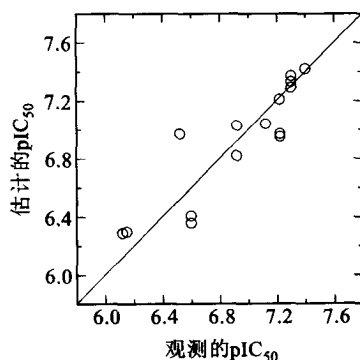
由方程 (6-1) 式估计的 pIC_{50} 值 (pIC_{BSR}) 及 LOO 检验预测的 pIC_{50} 值 (BSR_{LOO}) 与实验 pIC_{50} 值 (pIC_{50}) 一起列入表 6-3。估计 pIC_{50} 值与实验 pIC_{50} 值的整体相关情况如图 6-5 所示。由表 6-3 和图 6-5 可知, 除 7 号化合物的 pIC_{50} 估计值误差稍大之外, 其他化合物生物活性的估计误差较小。

表 6-3 16 个 ImAE 及其实验的、PLSR 模型估计的及 LOO 检验的 pIC_{50} 值

No	x_{14}	x_{22}	x_{29}	pIC_{50}	pIC_{BSR}	BSR_{LOO}
1	27.219 9	3.095 8	0.000 0	6.12	6.29	6.34
2	27.646 4	3.154 8	-0.012 3	6.15	6.30	6.33
3	27.547 7	3.723 1	0.000 0	6.60	6.36	6.31
4	28.257 8	3.204 5	0.106 4	6.60	6.41	6.37
5	46.510 7	3.331 3	0.099 9	6.92	6.82	6.81
6	46.523 4	10.330 3	0.000 0	7.40	7.42	7.46
7	48.439 9	5.138 0	0.000 0	6.52	6.97	7.03
8	49.813 1	4.701 9	0.000 0	7.22	6.95	6.91
9	44.355 5	6.806 0	0.000 0	7.12	7.04	7.02

续表

No	x_{14}	x_{22}	x_{29}	pIC_{50}	pIC_{BSR}	BSR_{LOO}
10	41.579 2	6.358 4	0.469 5	7.30	7.29	7.28
11	49.002 7	1.239 1	0.555 1	6.92	7.03	7.08
12	44.869 0	3.328 0	0.341 4	7.22	6.97	6.94
13	42.004 6	3.304 2	0.951 5	7.30	7.37	7.49
14	27.349 7	3.114 4	0.000 0	6.15	6.30	6.33
15	59.198 0	3.375 3	0.227 8	7.22	7.21	7.21
16	58.503 5	6.613 9	0.000 0	7.30	7.33	7.37

图 6-5 3 变量模型估计与实验 pIC_{50} 值相关图

在最佳子集回归模型中选择的 3 个 MEDV-13 描述子与 4 种原子类型 (第 2,3,6,10 原子类型) 相关联, 分别表示原子类型 2 与 2,2 与 10 以及 3 与 6 之间的相互作用。原子类型 2 与 3 主要存在于 ImAE 化合物的母体结构中, 而原子类型 10 在母体结构和取代基中都存在, 类型 6 则只存在于取代结构中。这说明 ImAE 化合物抑制 COX-2 的生物活性既与其母体结构有关也与取代基的结构有关。作为示例, 将 2 号 ImAE 分子中存在的原子类型与 MEDV 描述子标注在图 6-6 中。

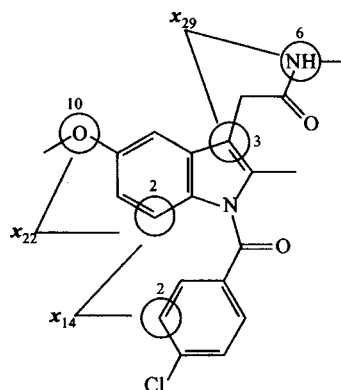


图 6-6 2号 ImAE 分子结构和原子类型及 MEDV描述子

6.1.3 PLSR 模型对应的似回归模型

偏最小二乘模型是一个关于主成分的隐含模型，根据第三章的方法原理可计算出关于原 33 个 MEDV 描述子的似 MLR 回归模型：

$$\begin{aligned}
 \text{pIC}_{50} = & 0.211\,04 - 0.336\,33\,x_1 + 0.003\,91\,x_2 - 0.008\,15\,x_3 - \\
 & 0.069\,54\,x_6 - 0.119\,61\,x_7 - 0.016\,41\,x_9 + \\
 & 0.018\,09\,x_{10} + 0.020\,98\,x_{13} + 0.006\,54\,x_{14} + \\
 & 0.006\,09\,x_{15} + 0.010\,01\,x_{18} - 0.205\,84\,x_{19} + \\
 & 0.017\,87\,x_{21} + 0.022\,73\,x_{22} + 0.002\,93\,x_{25} + \\
 & 0.065\,97\,x_{26} + 0.184\,94\,x_{29} - 0.168\,48\,x_{30} + \\
 & 0.014\,32\,x_{32} + 0.100\,24\,x_{33} - 0.020\,46\,x_{36} - \\
 & 0.440\,63\,x_{57} + 0.013\,44\,x_{59} + 0.229\,52\,x_{60} + \\
 & 0.609\,96\,x_{63} - 0.264\,85\,x_{66} - 0.390\,07\,x_{67} - \\
 & 0.202\,90\,x_{70} - 0.021\,13\,x_{77} + 0.019\,11\,x_{78} + \\
 & 0.017\,18\,x_{81} + 0.303\,13\,x_{82} + 0.076\,65\,x_{85}
 \end{aligned} \tag{6-2}$$

$m = 33, PC = 2, n = 16, r = 0.9514, RMSE = 0.14$ (建模)

$m = 33, PC = 2, n = 16, q = 0.8606, RMSV = 0.23$ (LOO 检验)

为了检验 PLSR 方法所定义的各个变量的 VIP 函数是否反映了各变量的重要程度, 我们选择 VIP 函数取值大于 0.80 的前 17 个 MEDV-13 描述子变量, 重新对 16 个 ImAE 化合物的 17 个 MEDV-13 做 PLSR 分析。结果发现, 这些 VIP 较大的描述子变量集的最佳主成分仍为 2 所得 PLSR 模型的估计与 LOO 检验统计量与原 33 个 MEDV 描述子集没有显著性差异, 这说明 PLSR 模型计算的 VIP 函数大小确实能够反映变量的重要性程度, 去掉某些次要因素后, 对模型结果没有显著性影响。

这个新的 PLSR 模型对应的似 MLR 回归模型如下:

$$\begin{aligned} pIC_{50} = & 2.22037 - 0.41825 x_1 - 0.06171 x_6 + 0.00713 x_{14} + \\ & 0.00172 x_{15} + 0.01601 x_{18} - 0.22065 x_{19} + \\ & 0.01652 x_{21} + 0.02320 x_{22} + 0.00989 x_{25} + \\ & 0.31757 x_{29} + 0.02590 x_{32} + 0.16249 x_{33} - \\ & 0.23161 x_{66} - 0.34401 x_{67} - 0.60468 x_{70} - \\ & 0.03084 x_{77} + 0.01480 x_{78} \end{aligned} \quad (6-3)$$

$m = 17, PC = 2, n = 16, r = 0.9467, RMSE = 0.14$ (建模)

$m = 17, PC = 2, n = 16, q = 0.8705, RMSV = 0.22$ (LOO 检验)

对于 ImAE 化合物, 我们曾应用主成分回归方法进行了分析建模^[33]。从 42 个非零 MEDV 描述子变量中, 首先删除非零变量数很少 (比如仅 1 个非零样本) 的描述子变量, 再删除显著相关的变量, 最后获得相关较小, 非零样本数较多的 32 个 MEDV-13 描述子。对由这 32 个描述子构建的数据集进行 PCR 分析可知最佳主成分数 $PC = 6$ 。尽管 PCR 模型估计与 LOO 交互检验结果的统计量还算满意, 对 4 个检验集的预测也是成功的。但对于只有 16 个化合物的 ImAE 化合物体系来说, 最佳主成分似嫌过大。结

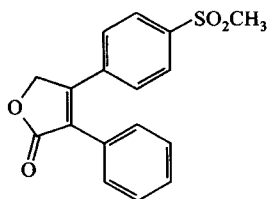
合我们上一章的 PCR 研究结果也会发现, 大多数情况下, 同一组数据集的 PCR 分析结果其主成分数都比 PLSR 结果要偏大一些, 这是由于 PCR 方法中只对分子描述子变量集进行主成分分解, 并没考虑对因变量如生物活性的相关所致。

6.2 二芳基环戊烯酮类 (DAP) 抑制剂

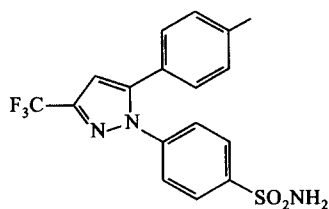
6.2.1 DAP 分子的结构表征

选择 Black 等^[34]研究过的一类二芳基环戊烯酮 (diarylcyclopentenone 简称 DAP) 类 COX-2 抑制剂为研究对象。DAP 类化合物包括 MK-0966、SC-58635 及 22 个二芳基环戊烯酮取代衍生物 (参见图 6-7) 共 24 个。这 24 个化合物抑制 COX-2 的实验生物活性取自该文献^[34]的 IC_{50} (μM) 值 并以其负对数 pIC_{50} 表示, $pIC_{50} = -\lg(IC_{50} \times 10^{-6})$ 。各个 DAP 分子的结构由 MEDV-13 描述子表征。由于在 DAP 分子中缺少某些原子类型, 因此实际上在 91 个 MEDV-13 分子描述子中只有 57 个具有 1 个或多个非零样本值, 其他 34 个 MEDV-13 描述子对 24 个 DAP 分子全部为零。在这 57 个非零描述子中, 有 21 个 MEDV 描述子 ($x_5, x_7, x_{17}, x_{19}, x_{28}, x_{30}, x_{38}, x_{39}, x_{40}, x_{46}, x_{48}, x_{49}, x_{51}, x_{54}, x_{55}, x_{57}, x_{60}, x_{66}, x_{69}, x_{70}, x_{85}$) 只有 1 个非零样本值, 另有 5 个描述子 ($x_4, x_{16}, x_{27}, x_{42}, x_{45}$) 只有 2 个非零样本值, 还有 1 个描述子 (x_1) 只有 3 个非零样本值 因此 在进行模型分析时 这 27 个统计显著性不明显的 MEDV 描述子首先从数据集中删除。

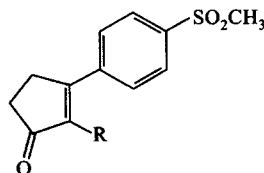
这 30 个 MEDV-13 描述子能完全地分辨 24 个 DAP 分子的结构 在这 30 个分子描述子变量空间中, 所有 DAP 分子对的模式距离都大于零, 其中最小样本模式距离为 $d(9,10)=0.5190$ 次最小距离为 $d(3,5)=0.8595$ 分辨率达 100%。



MK-0966



SC-58635



DAP母体结构

图 6-7 MK-0966, SC-58635 及 22 个二芳基环戊烯酮取代衍生物母体结构示意图

6.2.2 最佳主成分选择

以 24 个 DAP 化合物的 30 个非零 MEDV 描述子和 DAP 的 pIC_{50} 为数据集, 以 LOO 交互检验的均方根误差 $RMSV$ 为目标进行 PLSR 分析, 进而选择最佳主成分。PLSR 分析结果表明, 在建模阶段, 主成分 (PC) 增加, 均方根误差 $RMSE$ 下降。在 LOO 交互检验阶段, 均方根误差 $RMSV$ 先随主成分的增加而下降, 在 $PC=5$ 时达到最低点, 然后又随主成分的增加而上升, 在 $PC=12$ 时达到最高点 在 $PC=12$ 以后 $RMSV$ 出现不规则变化, 完整变化曲线如图 6-8 所示。由图可知, 最佳主成分数选 $PC=5$ 是合理的。在 $PC=5$ 时, PLSR 模型计算的 pIC_{50} 与实验 pIC_{50} 值的整体相关情况如图 6-9 所示。

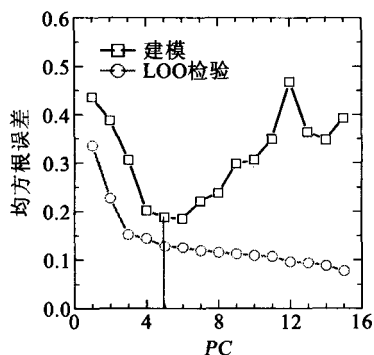


图 6-8 PLS 建模均方根误差随主成分的变化

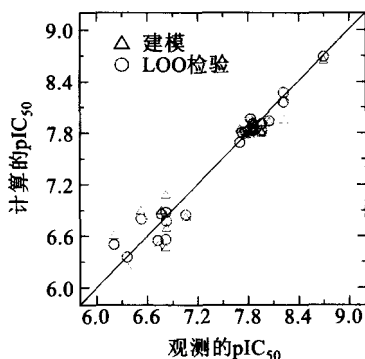


图 6-9 5 主成分 PLS 模型计算值与实验值相关图

在 $PC = 5$ 时 PLSR 模型的有关统计量分别为：

$n = 24, m = 30, PC = 5, r = 0.9816, RMSE = 0.128$ (建模)

$n = 24, m = 30, PC = 5, q = 0.9599, RMSV = 0.188$ (LOO 检验)

那么, 这 5 个主成分与原始 MEDV-13 分子描述子变量之间有何关系? 哪些 MEDV 描述子变量对 pIC_{50} 产生重要影响, 哪些只产生次要作用? 这可从 PLSR 分析中的 VIP 函数得出大致的

结论, VIP 函数取值越大, 相应的描述子对生物活性的影响就越大。PLSR 分析结果表明, VIP 取值按其从大到小的顺序排列, 前面 23 个重要 MEDV 描述子变量分别是:

$$\begin{aligned} & x_{18} (\text{VIP} = 1.889) \rightarrow x_{59} (1.743) \rightarrow x_{29} (1.689) \rightarrow x_{13} (1.394) \rightarrow \\ & x_{90} (1.382) \rightarrow x_{81} (1.331) \rightarrow x_6 (1.318) \rightarrow x_{24} (1.250) \rightarrow \\ & x_{25} (1.249) \rightarrow x_9 (1.195) \rightarrow x_{91} (1.165) \rightarrow x_3 (1.136) \rightarrow \\ & x_{12} (1.052) \rightarrow x_{77} (1.050) \rightarrow x_{80} (1.046) \rightarrow x_{14} (1.036) \rightarrow \\ & x_{63} (0.923) \rightarrow x_2 (0.909) \rightarrow x_{62} (0.868) \rightarrow x_{35} (0.818) \rightarrow \\ & x_{32} (0.802) \rightarrow x_{10} (0.783) \rightarrow x_{36} (0.780) \end{aligned}$$

为了分析影响 DAP 化合物 pIC_{50} 的最重要因素, 对上述 23 个 MEDV 描述子构成的数据集进行 BSR 分析, 获得不同变量数下的最佳变量组合模型, 各个模型的相关统计量结果见表 6-4。由表可知, 最佳模型为 $x_2, x_{18}, x_{24}, x_{29}$ 和 x_{36} 五个描述子构成的子集模型, 该模型估计相关系数 $r = 0.9768$ 和均方根误差 $\text{RMSE} = 0.143$, LOO 检验相关系数 $q = 0.9545$ 和均方根误差 $\text{RMSV} = 0.199$ 。显然, 模型具有良好的估计能力和稳定性。然而, BSR 选择的 5 变量模型比 PLSR 模型的质量稍差, 这是因为 PLSR 方法虽然只有 5 个主成分, 但利用了所有 30 个 MEDV 描述子表征的结构信息, 而最佳子集回归结果只利用了所选 5 个优化 MEDV 描述变量的结构信息, 丢失了其他描述子的结构信息。另一方面, 从 5 个优化 MEDV 描述变量之间的自相关系数可以看出, 除 x_{18} 和 x_{29} 间的相关系数 $r(18, 29) = 0.7799$ 有点偏大之外, 其余都小于 0.15 说明 BSR 得到的最佳子集的偶然相关较小。如果要求将进入最佳模型中的描述子之间的自相关系数控制在 0.15 以下, 那么最优模型为 MEDV-13 描述子 x_2, x_{24}, x_{36} 和 x_{59} 等构成的四变量子集模型, 该模型变量之间的最大自相关系数只有 $r(2, 36) = -0.1315$ 。

表 6-4 不同变量组合模型相关统计量和 MEDV 描述子组合

No	r^2	r	RMSE	q^2	q	RMSV	MEDV 描述子
1	0.582 2	0.763 0	0.432	0.539 1	0.734 3	0.454	x_{18}
2	0.919 6	0.959 0	0.190	0.829 1	0.910 6	0.279	x_9, x_{62}
3	0.924 7	0.961 6	0.183	0.870 2	0.932 9	0.242	x_9, x_{18}, x_{62}
4	0.936 2	0.967 6	0.169	0.885 0	0.940 8	0.235	$x_2, x_{24}, x_{36}, x_{59}$
5	0.954 1	0.976 8	0.143	0.911 2	0.954 5	0.199	$x_2, x_{18}, x_{24}, x_{29}, x_{36}$
6	0.957 7	0.978 6	0.137	0.911 8	0.954 9	0.200	$x_2, x_{18}, x_{24}, x_{29}, x_{35}, x_{36}$
7	0.961 9	0.980 8	0.130	0.914 5	0.956 3	0.196	$x_2, x_{14}, x_{18}, x_{29}, x_{36}, x_{63}, x_{90}$
8	0.962 0	0.980 8	0.130	0.910 2	0.954 0	0.202	$x_2, x_{14}, x_{18}, x_{29}, x_{35}, x_{36}, x_{63}, x_{90}$

6.2.3 QSAR 模型

由 5 个最佳主成分得到的 PLSR 模型经变换得到的似回归模型如下：

$$\begin{aligned}
 \text{pIC}_{50} = & 4.911\ 742 - 0.055\ 750\ x_2 - 0.049\ 920\ x_3 + \\
 & 0.020\ 228\ x_6 - 0.046\ 712\ x_9 + 0.065\ 212\ x_{10} + \\
 & 0.015\ 420\ x_{12} + 0.536\ 058\ x_{13} + 0.013\ 656\ x_{14} + \\
 & 0.002\ 746\ x_{15} - 0.022\ 197\ x_{18} + 0.048\ 285\ x_{21} + \\
 & 0.022\ 322\ x_{22} - 0.124\ 283\ x_{24} + 0.004\ 618\ x_{25} + \\
 & 0.011\ 292\ x_{26} - 0.056\ 161\ x_{29} + 0.010\ 439\ x_{32} + \\
 & 0.017\ 445\ x_{33} - 0.013\ 769\ x_{35} + 0.033\ 979\ x_{36} - \\
 & 0.174\ 560\ x_{59} + 0.338\ 669\ x_{62} + 0.010\ 872\ x_{63} + \\
 & 0.032\ 327\ x_{77} - 0.027\ 424\ x_{78} - 0.004\ 165\ x_{80} + \\
 & 0.055\ 607\ x_{81} - 0.130\ 580\ x_{84} - 0.191\ 594\ x_{90} + \\
 & 0.013\ 453\ x_{91}
 \end{aligned} \quad (6-4)$$

$m = 30, PC = 5, n = 24, r = 0.981\ 6, RMSE = 0.128$ (建模)

$m = 30, PC = 5, n = 24, q = 0.959\ 9, RMSV = 0.188$ (LOO 检验)

由 BSR 方法得到的 5 变量和 4 变量多元线性回归模型如下：

$$\begin{aligned} \text{pIC}_{50} = & (2.037\ 095 \pm 0.597\ 028) - (0.073\ 997 \pm 0.011\ 696)x_2 - \\ & (0.048\ 947 \pm 0.008\ 148)x_{18} - (0.633\ 509 \pm 0.063\ 357)x_{24} - \\ & (0.096\ 436 \pm 0.024\ 946)x_{29} + (0.034\ 788 \pm 0.016\ 412)x_{36} \end{aligned} \quad (6-5A)$$

$$m = 5, n = 24, r = 0.976\ 8, RMSE = 0.143 \quad (\text{建模})$$

$$m = 5, n = 24, q = 0.954\ 5, RMSV = 0.199 \quad (\text{LOO 检验})$$

$$\begin{aligned} \text{pIC}_{50} = & (0.973\ 169 \pm 0.682\ 215) - (0.075\ 749 \pm 0.013\ 408)x_2 - \\ & (0.744\ 576 \pm 0.072\ 879)x_{24} + (0.038\ 226 \pm 0.016\ 682)x_{36} - \\ & (0.709\ 896 \pm 0.055\ 377)x_{59} \end{aligned} \quad (6-5B)$$

$$m = 4, n = 24, r = 0.967\ 6, RMSE = 0.169 \quad (\text{建模})$$

$$m = 4, n = 24, q = 0.940\ 8, RMSV = 0.235 \quad (\text{LOO 检验})$$

由模型 (6-4), (6-5A) 和 (6-5B) 式估计的 DAP 生物活性 pIC_{50} (分别记为 PLS5, BSR5 与 BSR4) 以及实验 pIC_{50} (记为 pIC_{50}) 全部列入表 6-5 中。由模型 (6-4) 和 (6-5B) 式计算的 pIC_{50} 分别与实验的 pIC_{50} 的整体相关情况如图 6-10 和图 6-11 所示。

表 6-5 24 个 DAP 化合物及其实验的与 QSAR 模型计算的 pIC_{50} 数据

No	R	pIC_{50}	PLS5	BSR5	BSR4	PLS18	BSR18
1 (MK-0966)		7.699	7.688	7.805	7.777	7.785	7.807
2 (SC-58635)		8.699	8.689	8.700	8.702	8.722	8.701
3 4-fluorophenyl		7.721	7.806	7.844	7.823	7.779	7.797
4 phenyl		7.959	7.827	7.896	7.885	7.788	7.802
5 3-fluorophenyl		7.854	7.836	7.845	7.823	7.802	7.817
6 3,4-difluorophenyl		7.854	7.881	7.793	7.761	7.847	7.826

续表

No	R	pIC ₅₀	PLS5	BSR5	BSR4	PLS18	BSR18
7	3,5 - difluorophenyl	7.824	7.842	7.773	7.738	7.807	7.811
8	3,5 - dichlorophenyl	7.854	7.908	7.989	7.983	7.905	7.950
9	3 - chloro - 4 - fluorophenyl	8.046	7.94	7.937	7.923	7.915	7.939
10	3 - fluoro - 4 - chlorophenyl	7.824	7.963	7.943	7.928	7.937	7.952
11	3,4,5 - trichlorophenyl	8.222	8.157	8.155	8.160	8.125	8.193
12	2 - benzothiophenyl	7.854	7.824	7.975	7.978	7.953	7.781
13	3 - pyridyl	6.770	6.853	6.849	6.940	6.876	6.748
14	5 - chloro - 3 - pyridyl	7.060	6.847	6.817	6.938	6.892	6.779
15	5 - bromo - 3 - pyridyl	6.523	6.796	6.775	6.935	6.848	6.736
16	4 - methyl - 3 - pyridyl	6.357	6.355	6.500	6.610	6.415	6.558
17	4 - methoxy - 3 - pyridyl	6.824	6.876	6.612	6.736	6.780	6.587
18	2 - pyridyl	6.815	6.56	6.601	6.439	6.583	6.976
19	4 - bromo - 2 - pyridyl	6.201	6.502	6.530	6.443	6.555*	6.976*
20	4 - chloro - 2 - pyridyl	6.721	6.546	6.564	6.444	6.592*	7.009*
21	phenoxy	7.959	7.905	7.878	7.864	7.766*	7.795*
22	3,5 - difluorophenoxy	7.770	7.813	7.743	7.705	7.720*	7.765*
23	phenylacetylenyl	8.222	8.269	8.115	8.143	8.290*	7.813*
24	tert - butylacetylenyl	6.824	6.774	6.817	6.779	7.674*	7.423*

比较 PLSR 模型与 BSR 方法所得的 5 变量与 4 变量最佳子集模型可知,PLSR 模型质量明显优于 BSR 方法的 4 变量模型,稍优于 5 变量最佳子集模型。这是因为 PLSR 方法利用了所有的 30 个 MEDV 分子描述子,其分子结构信息比 4 或 5 变量表征的都得多得多,而 4 变量 MLR 模型只利用了最重要的 4 个描述子

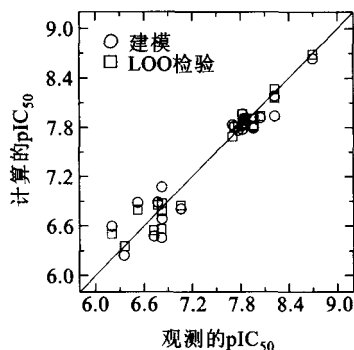


图 6-10 5 主成分 PLSR 模型计算与实验值相关图

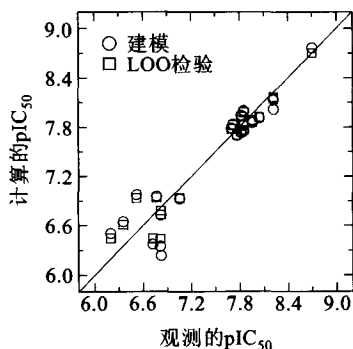


图 6-11 4 变量 BSR 模型计算与实验值相关图

对应的分子结构信息，因此统计结果稍差。5 个优化变量的 BSR 模型加入了更多的分子结构信息，其模型质量得到了提高。然而，正如前文所指出的，对于 24 个 DAP 样本数据集来说，5 变量模型不是非常满意的模型，同样 PLSR 模型包含 5 个主成分也似乎大了一些，这都需要加入更多的 DAP 化合物作为训练集得以改善。

对应 4 变量最佳子集模型的 4 个 MEDV-13 描述子是 DAP

分子中存在的 7 个原子类型 (第 1, 2, 3, 6, 9, 12 和 13 类型) 之间的 4 种相互作用的结果。这说明了 DAP 抑制剂母体结构中的子结构 $-\text{CH}_3(1)$ 、 $-\text{CH}=(2)$ 、 $>\text{C}=(3)$ 、 $=\text{O}(9)$ 和 $\geq\text{S}\leq(12)$ 以及取代基中的子结构 $-\text{N}=(6)$ 和 $-\text{F}, -\text{Cl}, -\text{Br}(13)$ 等是影响 DAP 化合物 pIC_{50} 的主要结构因素。那么这 4 个 MEDV 优化描述子能否完全地分辨 24 个 DAP 分子的结构变化? 计算 24 个 DAP 分子在 4 个 MEDV 描述子空间的模式距离可知, 两两样本分子对之间的模式距离都大于零, 其中最小模式距离是 4 号 DAP 和 21 号 DAP 之间的距离 $d(4, 21) = 0.0009$ 次小者为 1 号与 21 号 DAP 之间的距离 $d(6, 7) = 0.0172$ 。模式距离的结果充分说明了仅用 4 个优化 MEDV 描述子同样可完全分辨 DAP 的分子结构。

对应 5 变量最佳子集模型的 5 个 MEDV-13 描述子是 DAP 分子中存在的 7 个原子类型 (第 1, 2, 3, 6, 9, 12 和 13 类型) 之间的 5 种相互作用的结果。与 4 变量模型相比, 影响 pIC_{50} 的结构因素仍然是 DAP 母体结构中的子结构 $-\text{CH}_3(1)$ 、 $-\text{CH}=(2)$ 、 $>\text{C}=(3)$ 、 $=\text{O}(9)$ 和 $\geq\text{S}\leq(12)$ 以及取代基中的子结构 $-\text{N}=(6)$ 和 $-\text{F}, -\text{Cl}, -\text{Br}(13)$ 。这也说明 4 变量模型比 5 变量模型更优。

24 个 DAP 分子中与 BSR 模型相关的几个重要 MEDV-13 描述子值列入表 6-6 中。

表 6-6 24 个 DAP 化合物及其 7 个重要的 MEDV-13 描述子值

No	x_2	x_9	x_{18}	x_{24}	x_{29}	x_{36}	x_{59}
1	2.206 4	7.671 3	0.000 0	-9.362 0	0.000 0	0.000 0	0.000 0
2	4.563 5	0.264 2	8.301 1	-12.011 6	3.155 8	2.935 6	1.382 8
3	2.223 9	7.698 8	0.000 0	-9.432 3	0.000 0	-0.124 9	0.000 0
4	2.261 3	7.697 0	0.000 0	-9.512 7	0.000 0	0.000 0	0.000 0

续表

No	x_2	x_9	x_{18}	x_{24}	x_{29}	x_{36}	x_{59}
5	2.212 8	7.699 5	0.000 0	-9.403 9	0.000 0	0.397 9	0.000 0
6	2.168 0	7.701 4	0.000 0	-9.303 0	0.000 0	0.656 3	0.000 0
7	2.163 9	7.702 1	0.000 0	-9.293 0	0.000 0	0.254 0	0.000 0
8	2.221 0	7.756 8	0.000 0	-9.427 9	0.000 0	4.134 6	0.000 0
9	2.194 8	7.728 7	0.000 0	-9.366 2	0.000 0	3.709 1	0.000 0
10	2.191 5	7.724 2	0.000 0	-9.358 0	0.000 0	4.013 0	0.000 0
11	2.180 1	7.781 6	0.000 0	-9.331 2	0.000 0	10.572 3	0.000 0
12	2.306 1	7.809 8	0.000 0	-9.642 5	0.000 0	0.000 0	0.000 0
13	2.188 9	7.686 3	15.042 4	-9.323 6	2.039 1	0.000 0	1.139 9
14	2.170 2	7.716 2	15.398 5	-9.284 9	2.522 0	1.601 9	1.190 7
15	2.180 2	7.725 7	15.821 9	-9.308 4	2.786 2	1.307 7	1.202 4
16	5.299 6	8.620 2	15.297 6	-9.222 7	2.474 1	0.000 0	1.167 9
17	3.793 9	8.370 8	15.539 0	-9.251 7	2.536 7	0.000 0	1.180 4
18	2.173 5	7.683 7	11.803 1	-9.277 9	5.965 2	0.000 0	1.800 4
19	2.165 1	7.723 1	12.420 2	-9.263 7	6.777 5	1.332 3	1.852 3
20	2.155 0	7.713 6	12.149 7	-9.239 6	6.527 6	1.631 7	1.842 0
21	2.254 7	7.734 8	0.000 0	-9.483 9	0.000 0	0.000 0	0.000 0
22	2.172 0	7.738 5	0.000 0	-9.301 8	0.000 0	-0.761 1	0.000 0
23	2.385 0	7.635 4	0.000 0	-9.872 4	0.000 0	0.000 0	0.000 0
24	16.765 7	9.753 4	0.000 0	-9.502 7	0.000 0	0.000 0	0.000 0

为了进一步说明 MEDV-13 方法适用于小样本体系生物活性 QSAR 模型的构建和 MEDV 方法模型对外部样本的预测能力, 将 24 个 DAP 化合物分为两组进行检验。以前 18 个 DAP 化

合物作为训练集组，后 6 个 DAP 作为外部检验集组。首先，分别对训练集组进行 PLSR 分析和 BSR 变量优化得到最佳主成分和优化变量。PLSR 分析结果表明， $PC = 4$ 时的 PLSR 模型最优 相应的似 MLR 模型及有关统计量如下：

$$\begin{aligned}
 pIC_{50} = & 4.754\ 122 + 0.002\ 085\ x_2 - 0.047\ 565\ x_3 - \\
 & 0.305\ 659\ x_6 - 0.024\ 529\ x_9 + 0.045\ 955\ x_{10} + \\
 & 0.014\ 493\ x_{12} + 0.523\ 014\ x_{13} + 0.012\ 188\ x_{14} + \\
 & 0.007\ 671\ x_{15} - 0.019\ 845\ x_{18} + 0.063\ 377\ x_{21} + \\
 & 0.004\ 573\ x_{22} - 0.092\ 513\ x_{24} + 0.006\ 797\ x_{25} + \\
 & 0.011\ 268\ x_{26} - 0.051\ 483\ x_{29} - 0.004\ 777\ x_{32} - \\
 & 0.008\ 143\ x_{33} + 0.001\ 056\ x_{35} + 0.032\ 094\ x_{36} - \\
 & 0.161\ 658\ x_{59} + 0.177\ 201\ x_{62} + 0.019\ 004\ x_{63} + \\
 & 0.033\ 890\ x_{77} + 0.022\ 419\ x_{78} - 0.004\ 362\ x_{80} + \\
 & 0.051\ 188\ x_{81} - 0.294\ 092\ x_{84} - 0.186\ 021\ x_{90} + \\
 & 0.012\ 637\ x_{91}
 \end{aligned} \tag{6-6}$$

$$m = 30, PC = 4, n = 18, r = 0.978\ 8, RMSE = 0.129 \quad (\text{建模})$$

$$m = 30, PC = 4, n = 18, q = 0.937\ 2, RMSV = 0.224 \quad (\text{LOO 检验})$$

显然，由前 18 个训练集样本构建的模型具有良好相关性与稳定性，与 (6-4) 式由全部样本数据集建立的模型相比没有显著性差异。用该模型去预测 6 个检验集样本的生物活性，得到模型预测的 pIC_{50} 值与原实验 pIC_{50} 值之间的误差分别为：0.354, -0.129, -0.193, -0.050, 0.068 和 0.850。可以看出，最后一个检验集样本（即 24 号）的误差明显偏大，导致预测相关系数和均方根误差分别为： $r_p = 0.8739$, $RMSP = 0.380$ 。由这个训练集构建的 4 主成分模型估计训练集的 pIC_{50} 值 (PLS18) 和预测检验集的 pIC_{50} 值 (PLS18*) 的结果也列入前面的表 6-5 中。估计的与预测的 pIC_{50} 对实验的 pIC_{50} 的总体相关情况如图 6-12A 所示。

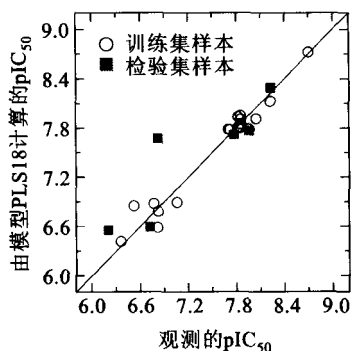


图 6-12A 18 个 DAP 训练集的 PLS 模型
计算与实验 pIC₅₀ 相关图

仿前节方法，求 4 个最佳主成分下各原始 MEDV-13 分子描述子变量的 VIP 函数，以便分析哪些 MEDV 描述子变量对 pIC₅₀ 产生重要影响，哪些只产生次要作用？VIP 函数取值越大，相应的分子描述子对生物活性的影响就可能越大。PLSR 分析结果表明，VIP 函数取值按从大到小的顺序排列时，前面 19 个 MEDV 描述子变量分别是（括号内数值为 VIP 值）：

$$\begin{aligned}
 & x_{18}(1.860) \rightarrow x_{59}(1.511) \rightarrow x_{13}(1.500) \rightarrow x_{90}(1.479) \rightarrow \\
 & x_{81}(1.476) \rightarrow x_{25}(1.420) \rightarrow x_{29}(1.389) \rightarrow x_6(1.284) \rightarrow \\
 & x_3(1.198) \rightarrow x_{24}(1.179) \rightarrow x_{91}(1.177) \rightarrow x_9(1.157) \rightarrow \\
 & x_{77}(1.060) \rightarrow x_{80}(1.052) \rightarrow x_{12}(1.051) \rightarrow x_{36}(1.019) \rightarrow \\
 & x_{63}(0.958) \rightarrow x_{10}(0.889) \rightarrow x_{14}(0.852)
 \end{aligned}$$

与应用全部 24 个 DAP 样本分子进行 PLSR 分析得到的重要描述子基本上是一样的，不同的只是 VIP 排位的顺序稍有变化。这也证明了 PLSR 方法中定义的 VIP 函数的有效性。

对 18 个 DAP 训练集化合物进行 BSR 分析的结果表明，3 个

MEDV 描述变量组合 (x_9, x_{18}, x_{36}) 为最佳子集, 其模型及有关统计量如下:

$$\begin{aligned} \text{pIC}_{50} = & (9.219\ 812 \pm 0.171\ 285) - (0.184\ 199 \pm 0.021\ 126)x_9 - \\ & (0.070\ 225 \pm 0.005\ 453)x_{18} + (0.038\ 500 \pm 0.014\ 593)x_{36} \end{aligned} \quad (6-7)$$

$$m = 3, n = 18, r = 0.976\ 3, \text{RMSE} = 0.136 \quad (\text{建模})$$

$$m = 3, n = 18, q = 0.963\ 7, \text{RMSV} = 0.170 \quad (\text{LOO 检验})$$

这说明 18 个训练集的 3 变量最佳子集模型具有良好的估计能力和稳定性。用这个模型去预测外部 6 个检验集样本, 预测误差分别为: 0.775, 0.288, -0.164, -0.005, -0.409, 0.599, 预测相关系数与均方根误差分别为: $r_p = 0.9458$ 和 $\text{RMSP} = 0.454$ 。尽管相关系数比 PLSR 模型还高, 但有三个样本的预测误差明显偏大。这是因为训练集样本较少, 得到的 3 个优化 MEDV 描述子表达的结构信息不能完全表征检验集样本的结构变化所致。要提高对这些样本的预测能力, 除了增加有代表性的训练集样本外, 适当增加有关分子描述子是可取的。

此外 尽管 BSR 选择的 3 变量 QSAR 模型具有与 PLS 同样的校正能力与稳定性, 但其对外部样本的预测能力明显不如 PLSR 方法, 这也说明了对于训练集样本较少的数据集, 最佳子集模型不如 PLSR 模型。这是因为 BSR 方法中选择的变量数太少 (比如 3 个) 对不能很好地反映 DAP 分子结构上的差异, 而 PLSR 方法则应用了更多的 MEDV-13 描述子的结构信息, 因而模型质量更高, 预测更准确。

为便于比较, 应用模型 (6-7) 式分别计算 18 个训练集与 6 个检验集的 pIC_{50} 的结果 (分别记为 BSR18) 也列入前面的表 6-5 中, 由模型 (6-7) 估计与预测的 18 个训练集与 6 个检验集样本的生物活性对原始观测值的相关情况如图 6-12B。3 个与 BSR 模型相关的 MEDV-13 分子描述子也列入前面的表 6-6 中。

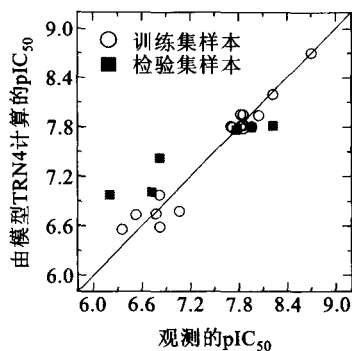


图 6-12B 18 个 DAP 训练集 3 描述子 BSR
模型计算与实验 pIC_{50} 相关图

6.3 噻唑酮与 噁唑酮类 (TOS) 抑制剂

6.3.1 TOS 分子的结构表征

选择 Song 等^[35]研究报道的噻唑酮与噁唑酮类 thiazolone and oxazolone series 简称 TOS) COX-2 抑制剂为研究对象。这些化合物包括吲哚美辛 (indomethacin)、SC-58125 和 NS-398 以及 TOS 衍生物 18 个 (骨架结构如图 6-13 所示), 一共 21 个化合物 统称为 TOS 化合物。这 21 个 TOS 化合物抑制 COX-2 的实验生物活性取自该文献^[35]的 IC_{50} (μM) 值, 并以负对数 $pIC_{50} = -\lg(IC_{50} \times 10^{-6})$ 表示。

各个 TOS 化合物的分子结构由 MEDV-13 描述子表征。在 21 个 TOS 分子中, 由于缺少第 8 和第 11 原子类型 因此 计算得到的 91 个 MEDV 描述子实际上有 29 个取值全部为零值, 即非零

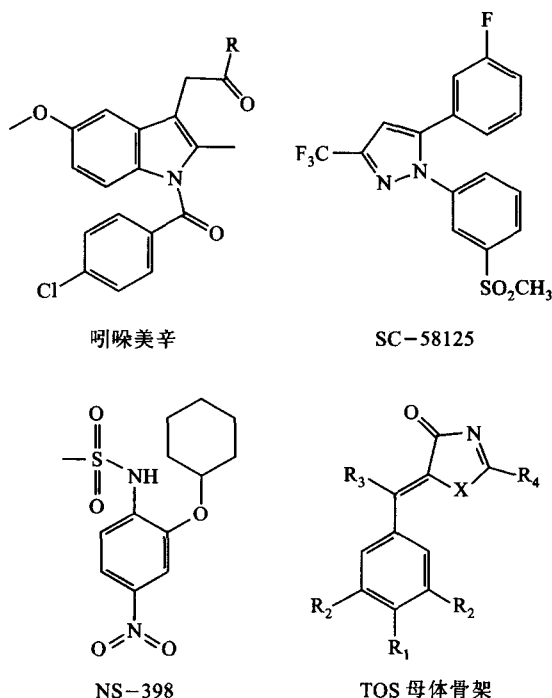


图 6-13 吲哚美辛、SC-58125、NS-398 以及
18 个 TOS 衍生物的骨架结构

MEDV-13 描述子只有 $91 - 29 = 62$ 个。这 62 个描述子能很好地分辨这 21 个 TOS 分子的结构，所有样本在描述子变量空间中是相互分离的，所有样本对的最小模式距离是第 2 号和第 3 号样本间的距离 $d(2,3) = 12.5951$ ，可见各个样本间的相似性很小。

由于各个 TOS 分子的原子类型不同，所含的分子描述子也不同。在 21 个 TOS 分子的 62 个非零 MEDV-13 描述子中，有 8 个描述子 ($x_{45}, x_{46}, x_{49}, x_{63}, x_{84}, x_{85}, x_{90}, x_{91}$) 只有 1 个样本具有非零值，另有 11 个描述子 ($x_{12}, x_{13}, x_{24}, x_{25}, x_{35}, x_{36}, x_{62}, x_{69}$,

x_{70}, x_{80}, x_{81})只有 2 个样本具有非零值, 还有 2 个描述子 (x_{67}, x_{40}) 只有 3 个样本非零值, 这 21 个 MEDV 描述子对生物活性的影响是不显著的, 不能充分反映 TOS 分子结构的变化, 可考虑从描述子集中删除。这样, 进入模型分析的 MEDV 描述子实际只余下 41 个, 21 个 TOS 化合物在这个新的 41 个描述子变量空间中也是相互分离的, 其最小模式距离还是 $d(2,3) = 12.5951$ 。

6.3.2 最佳主成分选择与 PLSR 分析

对上述选定的 41 个经初步选择的描述子集进行偏最小二乘回归分析, 结果发现比较合适的主成分为 $PC = 3$ 其 PLSR 似回归模型及相关统计量为:

$$\begin{aligned}
 \text{pIC}_{50} = & 5.683\,464 - 0.010\,617\,x_1 - 0.011\,572\,x_2 - \\
 & 0.010\,092\,x_3 + 0.003\,637\,x_4 - 0.085\,541\,x_5 - \\
 & 0.027\,576\,x_6 + 0.207\,852\,x_7 + 0.033\,887\,x_9 - \\
 & 0.040\,084\,x_{10} + 0.001\,691\,x_{14} - 0.009\,783\,x_{15} + \\
 & 0.024\,697\,x_{16} - 0.030\,793\,x_{17} + 0.028\,064\,x_{18} + \\
 & 0.051\,495\,x_{19} + 0.012\,997\,x_{21} - 0.012\,396\,x_{22} - \\
 & 0.007\,358\,x_{26} + 0.017\,781\,x_{27} - 0.010\,732\,x_{28} + \\
 & 0.043\,616\,x_{29} - 0.006\,552\,x_{30} + 0.015\,182\,x_{32} - \\
 & 0.055\,446\,x_{33} - 0.486\,412\,x_{37} + 0.296\,389\,x_{38} - \\
 & 0.045\,724\,x_{39} + 0.054\,535\,x_{42} + 0.388\,684\,x_{43} + \\
 & 0.004\,736\,x_{47} - 0.010\,750\,x_{48} - 0.050\,726\,x_{51} + \\
 & 0.310\,900\,x_{52} + 0.108\,431\,x_{56} - 0.098\,173\,x_{57} + \\
 & 0.013\,976\,x_{59} - 0.013\,718\,x_{60} + 0.229\,343\,x_{66} + \\
 & 0.011\,295\,x_{77} - 0.036\,170\,x_{78} - 0.432\,063\,x_{82} \quad (6-8)
 \end{aligned}$$

$m = 41, PC = 3, n = 21, r = 0.941\,1, RMSE = 0.25$ (建模)
 $m = 41, PC = 3, n = 21, q = 0.667\,7, RMSV = 0.56$ (LOO 检验)

由 6-8 式可知 建模相关系数与 LOO 检验相关系数相差较大 这可能是 21 个 TOS 化合物中描述子太多的缘故。为了寻找其原因,我们对 $PC=3$ 时的各原始 MEDV-13 变量的 VIP 函数进行分析研究。分析结果表明, VIP 函数值按从大到小顺序排列的前 24 个 MEDV-13 描述子为 (括号内为 VIP 值):

$x_{29}(1.601) \rightarrow x_{37}(1.511) \rightarrow x_1(1.496) \rightarrow x_4(1.489) \rightarrow$
 $x_{42}(1.400) \rightarrow x_{77}(1.383) \rightarrow x_{52}(1.372) \rightarrow x_{27}(1.325) \rightarrow$
 $x_{21}(1.323) \rightarrow x_3(1.264) \rightarrow x_2(1.234) \rightarrow x_{17}(1.228) \rightarrow$
 $x_{38}(1.225) \rightarrow x_{32}(1.214) \rightarrow x_{82}(1.209) \rightarrow x_{33}(1.208) \rightarrow$
 $x_{66}(1.184) \rightarrow x_{14}(1.183) \rightarrow x_5(1.173) \rightarrow x_{43}(1.173) \rightarrow$
 $x_{16}(1.024) \rightarrow x_{51}(0.996) \rightarrow x_{78}(0.987) \rightarrow x_{28}(0.980)$

取 VIP 值前 24 个 MEDV-13 描述子与 TOS 化合物的 pIC_{50} 值构成新的数据集,重新作 PLSR 分析 结果仍然是 $PC=3$ 为最佳,其相关统计量为:

$m=24, PC=3, n=21, r=0.9365, RMSE=0.26$ (建模)

$m=24, PC=3, n=21, q=0.7658, RMSV=0.48$ (LOO 检验)

显然与 41 个 MEDV 描述子构成的数据集相比,由 24 个 VIP 取值较大的重要 MEDV-13 描述子构成的数据集得到的 PLSR 模型 其 LOO 检验统计量更优。但与建模估计阶段的统计量相比仍有显著性差异,这可能是因为 21 个样本中存在某些离群值所致。我们进一步分析 $PC=3$ 的 QSAR 模型,考察每个样本的误差变化 发现第 6 号 TOS 的估计误差和 LOO 检验误差分别为 0.79 和 0.89 第 13 号 TOS 的估计与检验误差分别为 -0.34 和 -1.14。这两个化合物的误差都超过了其均方根误差的 2 倍,应认为是离群值。若从数据集中删除这两个化合物,再进行 PLSR 分析,可得 PLSR 模型的最佳主成分仍为 $PC=3$ 有关统计量为:

$m = 24, PC = 3, n = 19, r = 0.9836, RMSE = 0.14$ (建模)

$m = 24, PC = 3, n = 19, q = 0.8906, RMSV = 0.35$ (LOO 检验)

显然 无论建模估计或是 LOO 检验预测, 模型质量都好多了。

为了便于分析上述 PLSR 分析过程中均方根误差的变化规律, 将不同数量 TOS 样本与不同数量 MEDV 描述子数据集的建模估计均方根误差 ($RMSE$) 与 LOO 检验预测的均方根误差 ($RMSV$) 随主成分数 PC 的变化情况描述在图 6-14 中。

由图 6-14 可知, 原始 21 个 TOS 化合物和相应的 41 个 MEDV-13 描述子进行 PLSR 分析后得到的 $RMS-PC$ 曲线 (左图) 与经 VIP 函数选择的 24 个描述子变量中得到的 $RMS-PC$ 曲线 (中图) 类似, 但 LOO 检验预测的均方根误差明显降低, 与建模估计的均方根误差的差别也减小了。当去除两个离群样本后, LOO 检验均方根误差降低更加明显 (右图)。

对 19 个 TOS 化合物与 24 个 MEDV-13 描述子数据集进行 PLSR 分析 得到的 3 主成分似回归模型如下:

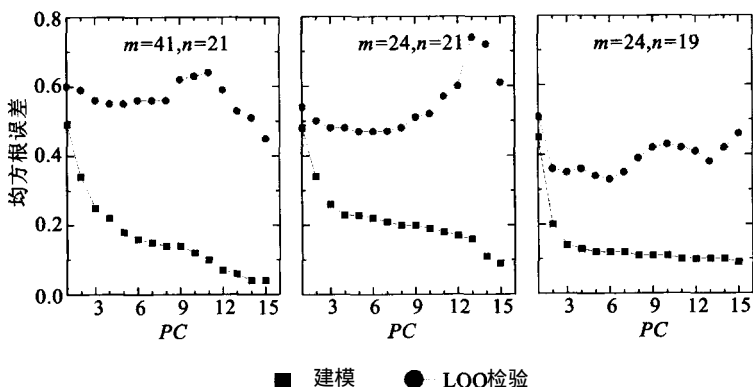


图 6-14 21 个 TOS 化合物体系 PLS 分析中 RMS 随主成分数的变化曲线
(左图: 21 个 TOS 41 个 MEDV 变量; 中图: 21 个 TOS 24 个 MEDV 变量;
右图: 19 个 TOS 24 个 MEDV 变量)

$$\begin{aligned}
pIC_{50} = & 5.360\ 931 - 0.010\ 560\ x_1 - 0.012\ 605\ x_2 - \\
& 0.005\ 697\ x_3 + 0.003\ 540\ x_4 - 0.102\ 436\ x_5 + \\
& 0.006\ 126\ x_{14} + 0.000\ 558\ x_{16} - 0.029\ 230\ x_{17} + \\
& 0.022\ 269\ x_{21} + 0.010\ 002\ x_{27} + 0.007\ 823\ x_{28} + \\
& 0.065\ 683\ x_{29} + 0.022\ 427\ x_{32} - 0.072\ 712\ x_{33} - \\
& 0.483\ 439\ x_{37} + 0.516\ 923\ x_{38} + 0.042\ 169\ x_{42} + \\
& 0.676\ 532\ x_{43} - 0.083\ 831\ x_{51} + 0.114\ 562\ x_{52} + \\
& 0.295\ 666\ x_{66} + 0.017\ 009\ x_{77} - 0.049\ 457\ x_{78} - \\
& 0.411\ 414\ x_{82} \quad (6-9)
\end{aligned}$$

$$m = 24, PC = 3, n = 19, r = 0.983\ 6, RMSE = 0.145 \quad (\text{建模})$$

$$m = 24, PC = 3, n = 19, q = 0.890\ 6, RMSV = 0.35 \quad (\text{LOO 检验})$$

为了进一步说明 MEDV-13 描述子模型的预测能力, 还须有检验集预测结果的证明。从 19 个 TOS 化合物中均匀选择 4 个化合物 No.3,9,14,19 为检验集, 其余 15 个化合物为训练集, 以训练集建立 PLSR 模型, 然后预测检验集。先对训练集进行主成分优化, 考察不同主成分下模型的统计性质, 结果表明 $PC = 3$ 时模型最佳, $PC = 3$ 时的似回归模型及有关统计结果如下:

$$\begin{aligned}
pIC_{50} = & 5.466\ 852 - 0.006\ 858\ x_1 - 0.007\ 963\ x_2 - \\
& 0.003\ 080\ x_3 + 0.002\ 637\ x_4 - 0.107\ 207\ x_5 - \\
& 0.003\ 322\ x_{14} + 0.023\ 194\ x_{16} - 0.026\ 578\ x_{17} + \\
& 0.018\ 534\ x_{21} + 0.010\ 478\ x_{27} + 0.000\ 782\ x_{28} + \\
& 0.064\ 423\ x_{29} + 0.017\ 656\ x_{32} - 0.061\ 100\ x_{33} - \\
& 0.375\ 247\ x_{37} + 0.490\ 312\ x_{38} + 0.038\ 518\ x_{42} + \\
& 0.671\ 460\ x_{43} - 0.109\ 121\ x_{51} + 0.273\ 460\ x_{52} + \\
& 0.290\ 524\ x_{66} + 0.015\ 525\ x_{77} - 0.042\ 765\ x_{78} - \\
& 0.504\ 091\ x_{82} \quad (6-10)
\end{aligned}$$

$$m = 24, PC = 3, n = 15, r = 0.982\ 8, RMSE = 0.14 \quad (\text{建模})$$

$m = 24$ $PC = 3$ $n = 15$ $q = 0.8568$ $RMSV = 0.38$ (LOO 检验)

上式统计结果说明该 PLSR 模型具有良好的估计能力和稳定性。利用这个模型去预测外部 4 个检验集样本, 结果表明预测均方根误差 $RMSP = 0.36$ 这 4 个检验样本 No. 3, 9, 14, 19 的预测误差分别为 0.24, -0.24, 0.15 和 -0.61。这个结果与训练集进行 LOO 检验的结果相当, 说明训练集模型具有良好预测能力。如果用这个模型去计算全部 19 个样本的 pIC_{50} 值, 则计算值与原实验值之间的相关系数为 0.9661 均方根误差为 0.20。

用模型 (6-9) 和 (6-10) 式计算的 19 个 TOS 样本的 pIC_{50} 值与实验 pIC_{50} 值的相关情况如图 6-15 所示。为了比较, 全部 21 个 TOS 化合物实验测定的 pIC_{50} (记为 pIC_{50})、利用 19 个 TOS 样本为训练集建立的 PLSR 模型计算的 pIC_{50} 值 (记为 PLS19) 和以 15 个 TOS 样本为训练集建立的 PLSR 模型计算的值 (记为 PLS15 全部列入表 6-7。

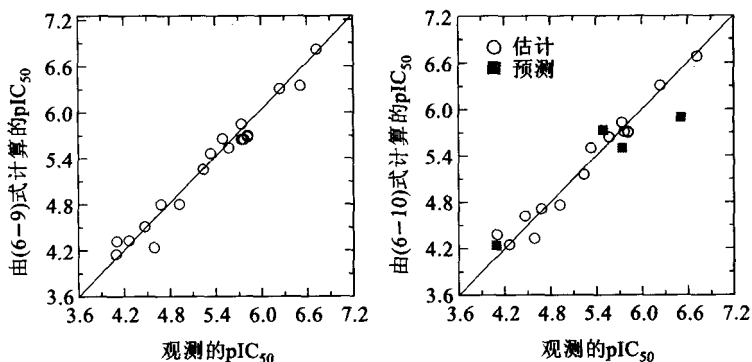


图 6-15 由模型(6-9)和(6-10)式计算的 19 个 TOS 样本的 pIC_{50} 值与实验 pIC_{50} 值相关图

表 6-7 21 个 TOS 抑制剂及其实验 pIC_{50} 与 PLSR 模型估计和预测 pIC_{50} 结果

No	cpd	X	R ₁	R ₂	R ₃	R ₄	IC ₅₀	pIC ₅₀	PLS19	PLS15
1	4a	S	OH	<i>t</i> -Bu	H	NHOH	1.5	5.82	5.68	5.70
2	4b	S	OH	<i>t</i> -Bu	H	NHOMe	1.7	5.77	5.65	5.71
3	4c	S	OH	<i>t</i> -Bu	H	NHOEt	3.2	5.49	5.66	5.73*
4	4d	S	OH	<i>t</i> -Bu	H	NHO-allyl	2.7	5.57	5.54	5.64
5	4e	S	OH	<i>t</i> -Bu	H	NHC(dNH)NH ₂	21	4.68	4.80	4.71
6	4q	S	OH	<i>t</i> -Bu	H	NMeOMe	18	4.74	5.63*	5.70*
7	4r	S	OH	<i>t</i> -Bu	H	NHCN	55	4.26	4.33	4.25
8	8ea	S	OH	<i>i</i> -Pr	H	NHOMe	0.57	6.24	6.31	6.31
9	8eb	S	OH	<i>i</i> -Pr	H	NHC(dNH)NH ₂	1.8	5.74	5.65	5.50*
10	9	S	OH	<i>t</i> -Bu	H	SMe	1.8	5.74	5.85	5.83
11	11c	O	OH	<i>t</i> -Bu	H	NHOEt	79	4.10	4.32	4.38
12	11d	O	OH	<i>t</i> -Bu	H	NHO-allyl	26	4.59	4.24	4.33
13	11e	O	OH	<i>t</i> -Bu	H	NHC(dNH)NH ₂	5.1	5.29	4.10*	4.39*
14	16d	NMe	OH	<i>t</i> -Bu	H	NHC(dNH)NH ₂	81	4.09	4.15	4.24*
15	17a	S	OH	<i>t</i> -Bu	H	OH	4.7	5.33	5.47	5.50
16	17b	O	OH	<i>t</i> -Bu	H	OH	34	4.47	4.51	4.62
17	18e	S	OH	<i>t</i> -Bu	Me	NHC(dNH)NH ₂	12	4.92	4.80	4.76
18	19a	S	OH	<i>t</i> -Bu	H	SH	1.5	5.82	5.70	5.71
19	SC-58125						0.31	6.51	6.35	5.90*
20	NS-398						0.19	6.72	6.82	6.68
21	indomethacin						5.8	5.24	5.26	5.16

注：数据后标有“*”的表示为模型对外部样本的预测值。

6.3.3 最佳子集回归与 MEDV 变量分析

前节 PLSR 分析获得了有关重要影响描述子的相关信息，并由原始 41 个 MEDV-13 描述子降维到 24 个，PLSR 统计性能没有发生变化。然而，24 个描述子进行定量结构活性相关仍然偏多，到底能不能应用更少的 MEDV 描述子变量来构建质量与 PLSR 模型匹配的 QSAR 模型 必须进行 BSR 分析。分别应用 21 个 TOS 与 41 个原始 MEDV 变量以及 21 个 TOS 与提取的 24 个 MEDV 变量构成的数据集，进行 BSR 优化变量分析，结果如表 6-8 所示。

表 6-8A 21 个 TOS 及其 41 个 MEDV 描述子数据集的 BSR 分析结果

m	r^2	r	RMSE	q^2	q	RMSV	MEDV 描述子
1	0.379 2	0.615 8	0.59	0.300 0	0.547 8	0.62	x_{37}
2	0.637 8	0.798 6	0.45	0.544 1	0.737 7	0.51	$x_1 x_{29}$
3	0.772 0	0.878 6	0.36	0.711 1	0.843 2	0.40	$x_1 x_{29} x_{52}$
4	0.844 9	0.919 2	0.29	0.784 2	0.885 5	0.35	$x_{33} x_{37} x_{52} x_{77}$
5	0.912 9	0.955 4	0.22	0.843 6	0.918 5	0.31	$x_3 x_6 x_{15} x_{29} x_{52}$
6	0.930 1	0.964 4	0.20	0.879 9	0.938 0	0.27	$x_1 x_6 x_{15} x_{29} x_{32} x_{52}$
7	0.944 3	0.971 7	0.18	0.911 3	0.954 6	0.22	$x_1 x_6 x_{15} x_{17} x_{29} x_{32} x_{52}$

表 6-8B 21 个 TOS 及其 24 个 MEDV 描述子数据集的 BSR 分析结果

m	r^2	r	RMSE	q^2	q	RMSV	MEDV 描述子
1	0.379 2	0.615 8	0.59	0.300 0	0.547 8	0.62	x_{37}
2	0.637 8	0.798 6	0.45	0.544 1	0.737 7	0.51	$x_1 x_{29}$
3	0.772 0	0.878 6	0.36	0.711 1	0.843 2	0.40	$x_1 x_{29} x_{52}$
4	0.844 9	0.919 2	0.29	0.784 2	0.885 5	0.35	$x_{33} x_{37} x_{52} x_{77}$
5	0.870 0	0.932 7	0.27	0.813 7	0.902 0	0.32	$x_5 x_{33} x_{37} x_{52} x_{77}$
6	0.874 5	0.935 1	0.26	0.800 9	0.895 0	0.33	$x_4 x_{17} x_{33} x_{38} x_{52} x_{77}$
7	0.881 9	0.939 1	0.26	0.792 6	0.890 3	0.34	$x_5 x_{14} x_{17} x_{33} x_{37} x_{52} x_{66}$
8	0.882 5	0.939 4	0.26	0.773 8	0.879 7	0.36	$x_4 x_5 x_{14} x_{17} x_{33} x_{52} x_{66} x_{77}$

表 6-8 结果表明 对于 21 个 TOS 与 41 个 MEDV 描述子构成的数据集, 均方根误差 $RMSE$ 与 $RMSV$ 都随变量数的增加而降低, 确定最佳变量组合比较困难。对于 21 个 TOS 与 24 个 MEDV 描述子构成的数据集, 最佳变量组合是 5 个 MEDV-13 描述子 ($x_5, x_{33}, x_{37}, x_{52}, x_{77}$) 构成的子集。考虑到数据集只有 21 个样本, 按经验规则其模型变量数不宜太多。通过比较两个数据集的 BSR 结果在 4 变量组合时选择的 MEDV 相同, 即都为描述子 $x_{33}, x_{37}, x_{52}, x_{77}$ (21 个 TOS 化合物的 4 个 MEDV 描述子数值参见表 6-10)。因此, 选择 4 变量子集模型更为合适。由这个子集建立的 MLR 模型如下:

$$pIC_{50} = (5.481\ 222 \pm 0.228\ 648) - (0.467\ 158 \pm 0.074\ 685)x_{33} - \\ (3.144\ 200 \pm 0.706\ 086)x_{37} + (0.872\ 823 \pm 0.197\ 297)x_{52} + \\ (0.060\ 268 \pm 0.014\ 448)x_{77} \quad (6-11)$$

$m=4, n=21, r=0.919\ 2, RMSE=0.29$ (建模)

$m=4, n=21, q=0.885\ 5, RMSV=0.35$ (LOO 检验)

与 3 个最佳主成分的 PLSR 模型相比, 该 4 描述子线性回归模型的建模估计统计量与 LOO 检验预测统计量结果和 PLSR 模型不存在显著性差异, 即模型具有良好估计能力和稳定性。参照 PLSR 方法, 从 21 个 TOS 数据集中去掉两个离群值 (第 6 和 13 号样本) 对 19 个 TOS 和 24 个 MEDV 描述子数据集进行 BSR 变量选择分析, 得到不同变量子集的各个 MLR 模型的估计与预测统计量如表 6-9 所示。

表 6-9 19 个 TOS 与 24 个 MEDV 描述子数据集的 BSR 分析结果

m	r^2	r	$RMSE$	q^2	q	$RMSV$	MEDV 描述子
1	0.376 7	0.613 7	0.61	0.287 1	0.535 8	0.65	x_{37}
2	0.760 6	0.872 1	0.38	0.703 6	0.838 8	0.42	x_{21}, x_{29}

续表

m	r^2	r	RMSE	q^2	q	RMSV	MEDV	描述子
3	0.854 8	0.924 6	0.29	0.806 1	0.897 8	0.34	x_{29}	x_{37} x_{77}
4	0.916 8	0.957 5	0.22	0.877 8	0.936 9	0.27	x_{29} x_{37}	x_{38} x_{77}
5	0.938 3	0.968 7	0.19	0.882 4	0.939 4	0.27	x_4 x_{29} x_{38}	x_{43} x_{77}
6	0.943 9	0.971 6	0.18	0.887 2	0.941 9	0.26	x_{17} x_{27} x_{29} x_{38}	x_{77} x_{78}
7	0.962 3	0.981 0	0.15	0.901 4	0.949 4	0.26	x_{16} x_{29} x_{32} x_{38}	x_{66} x_{77} x_{82}
8	0.964 7	0.982 2	0.15	0.899 3	0.948 3	0.29	x_{16} x_{17} x_{29} x_{32} x_{38}	x_{66} x_{77} x_{82}

从表中可以看出，最佳子集也是 4 个 MEDV - 13 描述子 (x_{29} 、 x_{37} 、 x_{38} 、 x_{77}) 构成的子集。由这 4 个描述子 (具体数值参见表 6-10) 构建的 MLR 方程及有关统计量如下：

$$\begin{aligned} \text{pIC}_{50} = & (4.982\ 538 \pm 0.193\ 945) + (0.195\ 346 \pm 0.025\ 179)x_{29} - \\ & (1.665\ 742 \pm 0.557\ 188)x_{37} + (1.199\ 371 \pm 0.371\ 554)x_{38} + \\ & (0.052\ 594 \pm 0.010\ 538)x_{77} \end{aligned} \quad (6-12)$$

$$m = 4, n = 19, r = 0.957\ 5, \text{RMSE} = 0.22 \quad (\text{建模})$$

$$m = 4, n = 19, q = 0.936\ 9, \text{RMSV} = 0.27 \quad (\text{LOO 检验})$$

分别由模型 6-11) 和 (6-12) 式计算的 21 个 TOS 的 pIC_{50} 值 (分别记为 MLR21 和 MLR19) 列入表 6-10 中。

为了说明 4 个描述子 MLR 模型的预测能力，以与 PLSR 方法选择检验集和训练集相同的方法进行建模检验，以 15 个 TOS 训练集样本进行 BSR 变量优化分析，可知 4 个 MEDV 描述子变量 (x_{29} 、 x_{52} 、 x_{77} 、 x_{82}) 组合构建的模型为优。应用 MLR 方法建立训练集模型，并用这个模型预测 4 个外部 TOS 样本 (NO. 3, 9, 14, 19) 的生物活性，其预测误差分别为：0.57、-1.14、-0.64、-0.32。可见第 9 号 TOS 化合物的预测误差明显偏大，这说明由 15 个 TOS 化合物与 4 个 MEDV 描述子构建的 MLR 模型预测能

表 6-10 21 个 TOS 抑制剂的 MEDV 描述子及其实验 pIC_{50} 与 MLR 模型估计和预测 pIC_{50} 结果

No	pIC_{50}	MLR21	MLR19	MLR15	x_{29}	x_{33}	x_{37}	x_{38}	x_{52}	x_{77}	x_{82}
1	5.82	5.50	5.52	5.62	5.064 2	-1.974 8	0.316 5	0.000 0	0.000 0	1.467 6	0.000 0
2	5.77	5.26	5.67	5.95	6.120 9	-1.603 2	0.316 5	0.000 0	0.000 0	0.401 1	-0.251 2
3	5.49	5.22	5.73	6.06*	6.443 2	-1.514 8	0.316 9	0.000 0	0.000 0	0.404 2	-0.312 4
4	5.57	5.23	5.62	5.96	5.898 4	-1.566 5	0.320 3	0.000 0	0.000 0	0.404 0	-0.308 0
5	4.68	4.69	4.63	4.60	3.504 4	-1.878 4	0.322 5	-0.430 7	-0.782 0	0.399 9	0.000 0
6	4.74	5.30	5.57*	5.89*	5.632 3	-1.694 5	0.318 3	0.000 0	0.000 0	0.405 4	-0.281 1
7	4.26	4.58	4.51	4.48	2.644 3	-1.623 6	0.325 8	-0.391 7	-0.758 6	0.392 2	0.000 0
8	6.24	6.23	6.17	5.92	5.992 4	-1.554 0	0.000 0	0.000 0	0.000 0	0.391 7	-0.245 0
9	5.74	5.68	5.66	4.60*	3.385 0	-1.795 5	0.000 0	0.000 0	-0.762 0	0.390 5	0.000 0
10	5.74	6.09	6.01	5.86	7.833 5	-3.352 8	0.312 7	0.000 0	0.000 0	0.401 0	0.201 2
11	4.10	4.05	4.69	4.42	1.118 2	0.955 4	0.322 2	0.000 0	0.000 0	0.397 8	0.539 6

续表

No	pIC ₅₀	MLR21	MLR19	MLR15	x ₂₉	x ₃₃	x ₃₇	x ₃₈	x ₅₂	x ₇₇	x ₈₂
12	4.59	4.16	4.60	4.37	0.686 7	0.695 0	0.325 7	0.000 0	0.000 0	0.397 6	0.501 3
13	5.29	5.50	3.69*	5.48*	-1.284 2	0.000 2	0.327 9	-0.430 3	1.178 2	0.393 5	0.000 0
14	4.09	4.47	3.99	4.73*	0.308 1	0.000 0	0.329 6	-0.436 3	0.000 0	0.401 9	0.000 0
15	5.33	5.23	5.00	5.17	2.213 7	-1.323 0	0.315 8	0.000 0	0.000 0	2.097 6	0.000 0
16	4.47	4.91	4.26	4.52	-1.514 1	-0.684 4	0.321 1	0.000 0	0.000 0	2.001 1	0.000 0
17	4.92	4.99	4.84	4.79	4.608 5	-2.554 7	0.325 2	-0.434 7	-0.792 4	0.409 3	0.000 0
18	5.82	5.50	5.57	5.65	5.324 8	-1.986 1	0.312 8	0.000 0	0.000 0	1.220 7	0.000 0
19	6.51	6.48	6.41	6.19*	2.833 8	0.000 0	0.000 0	0.000 0	0.000 0	16.599 1	0.000 0
20	6.72	6.70	6.76	6.74	0.704 4	1.425 8	0.000 0	0.000 0	0.000 0	31.206 7	0.000 0
21	5.24	5.39	5.45	5.22	0.000 0	1.345 1	0.000 0	0.000 0	0.000 0	8.898 4	0.000 0

力不是很理想，这主要是由于训练集样本太少，使得 4 个描述子表达的结构信息不能完全反映检验集中各个分子的结构信息的缘故。

由 15 个训练集构建的 4 变量（具体数值参见表 6-10）模型及统计量如下：

$$\begin{aligned} \text{pIC}_{50} = & (4.655\,452 \pm 0.144\,039) + (0.172\,416 \pm 0.029\,536)x_{29} + \\ & (0.869\,503 \pm 0.210\,658)x_{52} + (0.063\,038 \pm 0.009\,415)x_{77} - \\ & (0.847\,499 \pm 0.317\,506)x_{82} \quad (6-13) \\ m = 4, n = 15, r = 0.961\,7, RMSE = 0.20 & \quad (\text{建模}) \\ m = 4, n = 15, q = 0.909\,5, RMSV = 0.33 & \quad (\text{LOO 检验}) \end{aligned}$$

若采用 19 个 TOS 样本经 BSR 分析得到的 4 个最佳子集变量 ($x_{29}, x_{37}, x_{38}, x_{77}$) 来建立 15 个 TOS 训练集模型，则模型如下：

$$\begin{aligned} \text{pIC}_{50} = & (4.864\,428 \pm 0.272\,707) + (0.206\,108 \pm 0.030\,316)x_{29} - \\ & (1.338\,553 \pm 0.784\,972)x_{37} + (1.347\,497 \pm 0.459\,085)x_{38} + \\ & (0.055\,203 \pm 0.013\,513)x_{77} \quad (6-14) \\ m = 4, n = 15, r = 0.948\,1, RMSE = 0.23 & \quad (\text{建模}) \\ m = 4, n = 15, q = 0.902\,5, RMSV = 0.33 & \quad (\text{LOO 检验}) \end{aligned}$$

由上述研究结果可知，与 21 个及 19 个 TOS 样本数据集构建的模型相比，15 个 TOS 训练集构建的模型（6-13）式中选择的最佳变量是不相同的，这也说明了 15 个训练集样本太少，使得最佳子集中分子描述子发生了变化。由（6-13）式计算的 21 个 TOS 的 pIC_{50} 值（记为 MLR15）也列入表 6-10 中。为了观察计算结果的总体相关情况，将模型（6-12）和（6-13）式计算的 pIC_{50} 值分别对实验观测的 pIC_{50} 值作图如图 6-16 所示。

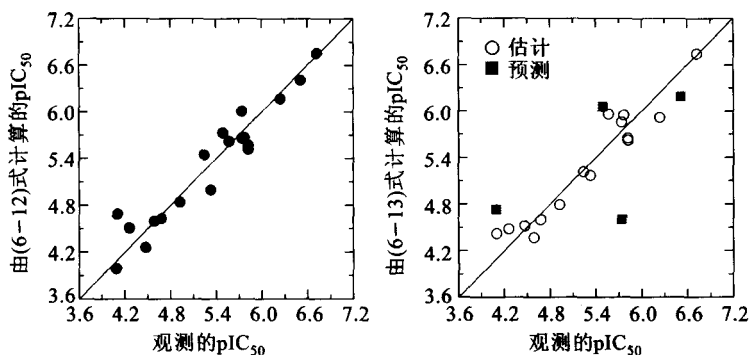


图 6-16 由模型 (6-12) 和 (6-13) 式计算的 19 个 TOS 样本的 pIC_{50} 值与实验 pIC_{50} 值相关图

6.4 3,4-二芳基噁唑酮 (DAA) 类抑制剂

6.4.1 DAA 分子结构表征

选择 Puig 等^[36]研究过的 3,4-二芳基噁唑酮 (diaryloxazolones 简称 DAA) 类 COX-2 抑制剂及其生物活性数据为数据集。该数据集包括母体结构如图 6-17 的 3,4-二芳基噁唑酮类取代衍生物 31 种, 加上吲哚美辛 (indomethacin 结构见图 6-13)、萘普生 (naproxen 和 celecoxib 图 6-17) 三种抑制剂, 总共 34 个化合物 (原文献中还包含 2 个化合物, 因为只有 IC_{50} 的大致范围, 这里没有选入)。为了方便, 全部统称为 DAA 化合物。这 34 个化合物抑制 COX-2 的实验生物活性取自该文献^[36]的 IC_{50} (μM) 值, 并以其负对数 pIC_{50} 表示。

各个 DAA 化合物分子结构由 MEDV-13 描述子表征。在所

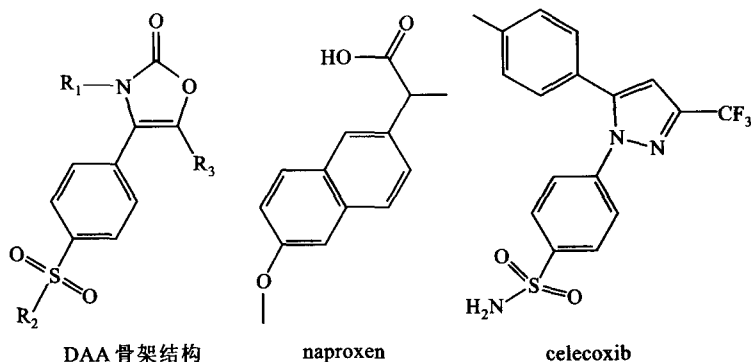


图 6-17 Naproxen, Celecoxib 和 DAA 骨架结构

有 34 个 DAA 分子中，由于不存在某些原子类型（如第 6, 8 和第 11 原子类型）致使 91 个 MEDV-13 描述子中有 31 个都没有非零样本值，因而实际上非零 MEDV 描述子只有 60 个。它们的编号与原子类型编号之间的关系见表 6-11。这 60 个 MEDV 描述子能很好地分辨这 34 个样本分子的结构，34 个 DAA 在 MEDV 描述子变量空间中是相互分离的，即任意两样本间的模式距离都大于零，其中最小距离为第 17 号和第 18 号样本间的距离 $d(17, 18) = 0.7786$ 从这两个 DAA 的结构来看是非常相似的，只是甲基取代基的位置稍有不同。

表 6-11 34 个 DAA 化合物和实验生物活性及模型计算生物活性

No	R ₁	R ₂	R ₃	IC ₅₀	pIC ₅₀	PLS30	PLS24
1	C ₆ H ₅	CH ₃	H	1.60	5.80	5.97	6.00
2	2-FC ₆ H ₄	CH ₃	H	0.66	6.18	6.02	6.07
3	4-FC ₆ H ₄	CH ₃	H	0.51	6.29	6.13	6.07
4	4-ClC ₆ H ₄	CH ₃	H	0.32	6.49	6.26	6.24

续表

No	R ₁	R ₂	R ₃	IC ₅₀	pIC ₅₀	PLS30	PLS24
5	4 - CF ₃ C ₆ H ₄	CH ₃	H	2.00	5.70	5.85	5.90
6	4 - CH ₃ C ₆ H ₄	CH ₃	H	0.48	6.32	6.30	6.31
7	4 - CH ₃ CH ₂ C ₆ H ₄	CH ₃	H	0.69	6.16	6.16	6.17
8	2,4 - diFC ₆ H ₃	CH ₃	H	1.35	5.87	6.07	6.03
9	C ₆ H ₅	NH ₂	H	2.40	5.62	5.62	5.64
10	2 - FC ₆ H ₄	NH ₂	H	2.20	5.66	5.67	5.70
11	4 - FC ₆ H ₄	NH ₂	H	1.50	5.82	5.78	5.70
12	2 - ClC ₆ H ₄	NH ₂	H	6.50	5.19	5.70	5.73
13	3 - ClC ₆ H ₄	NH ₂	H	1.50	5.82	5.89	5.86
14	4 - ClC ₆ H ₄	NH ₂	H	1.20	5.92	5.91	5.87
15	4 - CF ₃ C ₆ H ₄	NH ₂	H	2.00	5.70	5.50	5.53
16	2 - CH ₃ C ₆ H ₄	NH ₂	H	9.70	5.01	6.14 [*]	6.17
17	3 - CH ₃ C ₆ H ₄	NH ₂	H	0.51	6.29	6.20	6.16
18	4 - CH ₃ C ₆ H ₄	NH ₂	H	0.79	6.10	6.16	6.10
19	4 - CH ₃ CH ₂ C ₆ H ₄	NH ₂	H	1.56	5.81	6.00	5.95
20	4 - CH ₃ OC ₆ H ₄	NH ₂	H	0.21	6.68	6.00 [*]	5.97
21	2,4 - diFC ₆ H ₃	NH ₂	H	2.30	5.64	5.72	5.66
22	3,4 - diClC ₆ H ₃	NH ₂	H	0.40	6.40	6.26	6.24
23	4 - CH ₃ O, 3 - FC ₆ H ₃	NH ₂	H	1.13	5.95	5.91	5.99
24	4 - CH ₃ O, 3 - ClC ₆ H ₃	NH ₂	H	0.78	6.11	6.13	6.14
25	cyclohexyl	NH ₂	H	18.90	4.72	4.65	4.69
26	1 - naphthyl	NH ₂	H	4.70	5.33	5.13	5.12
27	C ₆ H ₅	NH ₂	CH ₃	1.80	5.74	5.93	6.14 [*]

续表

No	R ₁	R ₂	R ₃	IC ₅₀	pIC ₅₀	PLS30	PLS24
28	4-FC ₆ H ₄	NH ₂	CH ₃	0.51	6.29	6.08	6.21*
29	3-CH ₃ C ₆ H ₄	NH ₂	CH ₃	0.93	6.03	5.85	6.38*
30	4-CH ₃ C ₆ H ₄	NH ₂	CH ₃	1.70	5.77	5.97	6.39*
31	3,4-diClC ₆ H ₃	NH ₂	CH ₃	0.23	6.64	6.57	6.75*
32	indomethacin			0.22	6.66	7.76*	10.91*
33	naproxen			18.90	4.72	6.89*	9.87*
34	celecoxib			1.10	5.96	5.93	5.76*

然而,60 个描述子中有多数 MEDV 描述子的非零样本数很少,显著性不充分,应首先删除。经分析可知,60 个非零 MEDV 描述子中有 9 个描述子 ($x_6, x_{18}, x_{29}, x_{39}, x_{48}, x_{57}, x_{59}, x_{62}, x_{63}$) 只有 1 个样本取非零值,2 个描述子 (x_4, x_{38}) 只有 2 个样本为非零值,8 个描述子 ($x_{16}, x_{27}, x_{40}, x_{42}, x_{43}, x_{45}, x_{46}, x_{82}$) 只有 3 个样本为非零值,从变量统计性考虑首先删除这 19 个描述子,这样还余下 41 个描述子。那么这 41 个 MEDV 描述子能否完全分辨 34 个 DAA 分子的结构?这可以通过计算 DAA 分子两两之间在这些描述子变量空间中的模式距离加以解答。计算结果表明,所有模式距离都大于零,说明这 34 个 DAA 分子是相互分离的,其中最小距离仍是第 17 号和第 18 号样本间的距离,即 $d(17,18)=0.7786$ 。

6.4.2 主成分选择与 PLSR 模型

以 34 个 DAA 化合物的 41 个非零 MEDV-13 描述子和生物活性 pIC₅₀ 为数据集进行 PLSR 分析,结果建模估计均方根误差 RMSE 和 LOO 检验预测均方根误差 RMSV 随主成分 PC 的变化情况如图 6-18 所示。由图可知,在建模估计阶段, RMSE 随 PC 的增加而逐渐降低。然而,在 LOO 检验预测阶段, RMSV 随 PC

的变化规律性不强,存在多个极值点 ($PC = 2, 16, 19$ 等),这说明这个数据集中描述子变量不能很好地说明生物活性的变化规律,可能是样本集中存在某些离群值。显然在 3 个极值点中, $PC = 16$ 和 19 两个太大,应取 $PC = 2$ 进行样本误差分析。在 $PC = 2$ 时,PLSR 模型统计量如下:

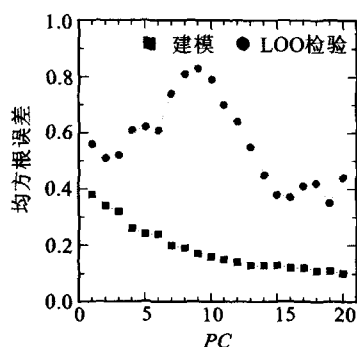


图 6-18 34 个 DAA 与 41 个 MEDV 描述子模型
均方根误差随 PC 的变化

$m = 41, PC = 2, n = 34, r = 0.7055, RMSE = 0.34$ (建模)

$m = 41, PC = 2, n = 34, q = 0.1765, RMSV = 0.51$ (LOO 检验)

应用这个模型进行 LOO 检验分析可知,在预测得到的 34 个 DAA 分子生物活性中,第 16, 20, 32 和 33 号 DAA 化合物的误差超过了均方根误差的 2 倍以上,分别达到 1.02, -0.91, -1.16 和 1.30。除去这 4 个 DAA,以 30 个 DAA 化合物集重建 PLSR 模型,结果均方根误差随 PC 的变化如图 6-19 所示。

显然,均方根误差明显低于 34 个 DAA 集构建的模型,最佳主成分数 $PC = 7$,最佳模型的有关统计量如下:

$m = 41, PC = 7, n = 30, r = 0.9331, RMSE = 0.14$ (建模)

$m = 41, PC = 7, n = 30, q = 0.8142, RMSV = 0.24$ (LOO 检验)

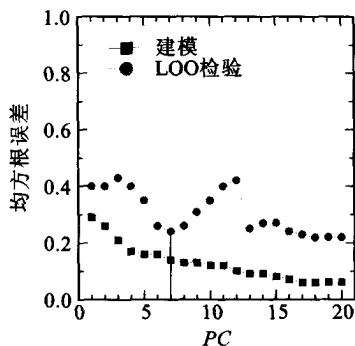


图 6-19 30 个 DAA 与 41 个 MEDV 描述子模型
均方根误差随 PC 的变化

分析 7 主成分 PLSR 模型，可知影响 DAA 化合物生物活性的重要 MEDV-13 描述子 如果按 VIP 函数大小排序 则 VIP 值大于 1.000 的 MEDV 描述子 (括号内为 VIP 值) 是：

$x_1(2.004) \rightarrow x_{36}(1.981) \rightarrow x_{14}(1.965) \rightarrow x_{19}(1.576) \rightarrow$
 $x_{67}(1.547) \rightarrow x_{30}(1.537) \rightarrow x_3(1.520) \rightarrow x_{66}(1.505) \rightarrow$
 $x_{69}(1.441) \rightarrow x_{91}(1.435) \rightarrow x_2(1.420) \rightarrow x_7(1.394) \rightarrow$
 $x_9(1.360) \rightarrow x_5(1.296) \rightarrow x_{80}(1.271) \rightarrow x_{17}(1.267) \rightarrow$
 $x_{28}(1.260) \rightarrow x_{12}(1.260) \rightarrow x_{51}(1.249) \rightarrow x_{54}(1.247) \rightarrow$
 $x_{10}(1.221) \rightarrow x_{24}(1.174) \rightarrow x_{52}(1.161) \rightarrow x_{26}(1.113) \rightarrow$
 $x_{35}(1.053) \rightarrow x_{70}(1.050) \rightarrow x_{49}(1.045) \rightarrow x_{85}(1.013)$

应用这 28 个重要 MEDV 描述子与 30 个 DAA 构建数据集，重新做 PLSR 分析，其建模估计均方根误差与 LOO 检验预测均方根误差随 PC 的变化如图 6-20 所示。与图 6-19 相比较，最佳主成分由 7 变为 6，而模型质量并没有显著性改变，说明将 41 维 MEDV 描述子降低为 28 维是合理的，删除的描述子确实对生物活性的贡献不大。该 PLSR 模型及对应的相关统计量如下：

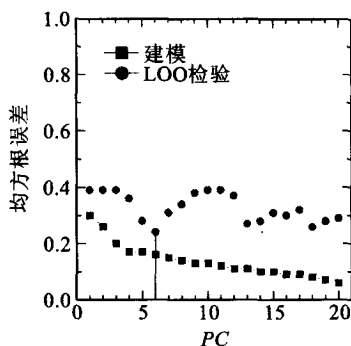


图 6-20 30 个 DAA 与 28 个 MEDV 描述子模型
均方根误差随 PC 的变化

$$\begin{aligned}
 \text{pIC}_{50} = & 8.388\,473 - 5.135\,150\,x_1 + 0.020\,395\,x_2 + \\
 & 0.104\,327\,x_3 + 0.683\,499\,x_5 - 1.032\,149\,x_7 + \\
 & 0.006\,577\,x_9 - 0.003\,276\,x_{10} - 0.002\,297\,x_{12} - \\
 & 0.030\,072\,x_{14} - 0.012\,709\,x_{17} - 0.301\,673\,x_{19} - \\
 & 0.008\,365\,x_{24} - 0.092\,332\,x_{26} - 0.034\,470\,x_{28} - \\
 & 0.069\,173\,x_{30} + 0.028\,530\,x_{35} + 0.097\,659\,x_{36} - \\
 & 2.409\,220\,x_{49} - 0.002\,371\,x_{51} - 0.559\,826\,x_{52} + \\
 & 0.001\,430\,x_{54} - 0.122\,462\,x_{66} - 0.647\,007\,x_{67} + \\
 & 1.704\,505\,x_{69} + 0.638\,672\,x_{70} + 0.019\,187\,x_{80} - \\
 & 0.372\,308\,x_{85} - 0.026\,493\,x_{91} \quad (6-15)
 \end{aligned}$$

$$m = 28, PC = 6, n = 30, r = 0.911\,9, RMSE = 0.16 \quad (\text{建模})$$

$$m = 28, PC = 6, n = 30, q = 0.809\,2, RMSV = 0.24 \quad (\text{LOO 检验})$$

应用模型 (6-15) 计算的 34 个 DAA 化合物的 pIC_{50} (记为 PLS_{30}) 与实验 pIC_{50} 值 (pIC_{50}) 列入表 6-11, 其整体相关情况如图 6-21。

为了说明 PLSR 方法所建模型对外部样本的预测能力, 选择

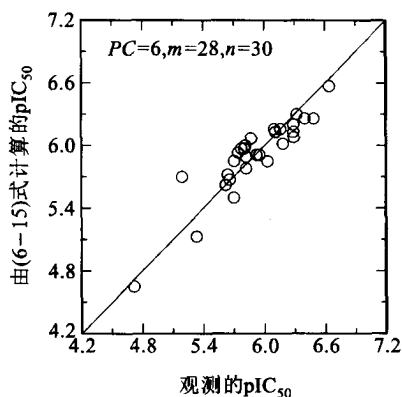


图 6-21 (6-15)式计算的与实验的 pIC_{50} 相关图

30个样本中的前24个DAA化合物的28个MEDV-13描述子与相应的 pIC_{50} 构成训练集后6个DAA为检验集,以训练集建立PLSR模型。结果表明,4主成分PLSR模型最优其似MLR模型及其统计量如下:

$$\begin{aligned}
 pIC_{50} = & 10.899\ 870 - 2.221\ 175\ x_1 + 0.013\ 116\ x_2 + \\
 & 0.074\ 819\ x_3 + 1.739\ 956\ x_5 - 1.618\ 214\ x_7 + \\
 & 0.003\ 731\ x_9 - 0.002\ 403\ x_{10} - 0.002\ 155\ x_{12} - \\
 & 0.020\ 220\ x_{14} - 0.012\ 452\ x_{17} - 0.216\ 372\ x_{19} - \\
 & 0.114\ 364\ x_{24} - 0.101\ 061\ x_{26} - 0.031\ 917\ x_{28} - \\
 & 0.064\ 013\ x_{30} + 0.126\ 407\ x_{35} + 0.120\ 960\ x_{36} - \\
 & 1.403\ 545\ x_{49} - 0.002\ 493\ x_{51} - 0.398\ 518\ x_{52} + \\
 & 0.001\ 522\ x_{54} - 0.117\ 898\ x_{66} - 0.724\ 620\ x_{67} + \\
 & 1.586\ 228\ x_{69} + 0.663\ 711\ x_{70} + 0.024\ 226\ x_{80} - \\
 & 0.235\ 398\ x_{85} - 0.025\ 180\ x_{91}
 \end{aligned} \tag{6-16}$$

$m = 28, PC = 4, n = 24, r = 0.906\ 9, RMSE = 0.17$ (建模)

$m = 28$ $PC = 4$ $n = 24$ $q = 0.5452$ $RMSV = 0.34$ (LOO 检验)

显然 ,LOO 检验结果较 6-15 式稍差。用该模型预测外部的 6 个检验集样本的 pIC_{50} 除 30 号 DAA 的误差较大 (0.62) 外 , 其余都较小。第 27,28,29,30,31 和 34 号等 6 个 DAA 化合物的误差分别为 0.40, -0.08, 0.35, 0.62, 0.11 和 -0.20 说明模型预测效果良好。由 (6-16) 式计算的 pIC_{50} 值 (记为 PLS24) 也列入表 6-11。其与实验 pIC_{50} 值的整体相关情况见图 6-22。

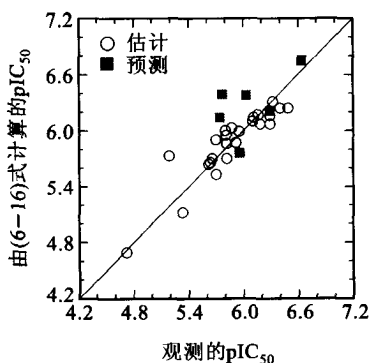


图 6-22 (6-16) 式计算的与实验的 pIC_{50} 相关图

参 考 文 献

- [1] Hla T, Neilson K. Human cyclooxygenase - 2 cDNA. Proc. Natl. Acad. Sci. U. S. A., 1992, 89: 7384~7388
- [2] Fu J Y, Masferrer J L, Siebert K, et al. The induction and suppression of prostaglandin H_2 synthase (COX) in human monocytes. J Biol. Chem, 1990, 265: 16737~16740
- [3] Xie W, Chipman J G, Robertson D L, Erikson R L, Simmons D L. Expression of a mitogen - responsive gene encoding prostaglandin synthase is regu-

- lated by mRNA splicing. *Proc. Natl. Acad. Sci. U.S.A.*, 1991, 88: 1692 ~ 1696
- [4] Wilkerson W W, Copeland R A, Covington M B, et al. Antiinflammatory 4, 5 - diarylpyrroles. 2. activity as a function of cyclooxygenase - 2 inhibition. *J. Med. Chem.*, 1995, 38(20): 3895 ~ 3901
- [5] Vane J R. Towards a better aspirin. *Nature Struct. Biol.*, 1994, 36(2): 215 ~ 216
- [6] Groosman C G, Wiseman J, Luess F S, et al. Inhibition of COX - 1 and COX - 2 in human platelets and mononuclear cells by NSAIDs. *Inflamm. Res.*, 1995, 44(suppl. 3): S272
- [7] 赵丽琴, 张守芳. 环氧合酶 - 2 选择性抑制研究进展. *中国药物化学杂志*, 1999, 9(2): 138 ~ 144
- [8] Herchsmann H R. Prostaglandin synthase 2. *Biochim. Biophys. Acta*, 1996, 1299: 125 ~ 140
- [9] O'Banion M K, Winn V D, Young D A. cDNA cloning and functional activity of a glucocorticoid - regulated inflammatory cyclooxygenase. *Proc. Natl. Acad. Sci. U.S.A.*, 1992, 89: 4888 ~ 4892
- [10] Puig C, Crespo M I, Godessart N, et al. Synthesis and biological evaluation of 3, 4 - diaryloxazolones: a new class of orally active cyclooxygenase - 2 inhibitors. *J. Med. Chem.*, 2000, 43(2): 214 ~ 223
- [11] Black W C, Brideau C, Chan C C, et al. 2, 3 - Diarylcyclopentenones as orally active, highly selective COX - 2 inhibitors. *J. Med. Chem.*, 1999, 42(7): 1274 ~ 1281
- [12] Futaki N, Takahashi S, Yoshikawa K, Arai I, Higuchi S, Otomo S. NS - 398, a new antiinflammatory agent, selectively inhibits prostaglandin G/H synthase/COX - 2 activity in vitro. *Prostaglandins*, 1994, 47: 55 ~ 59
- [13] Kalgutkar A S, Kozak K R, Crews B C, Hochgesang J G P, Marnett L J. Covalent modification of cyclooxygenase - 2 (COX - 2) by α - acetoxyphe-nyl alkyl sulfides, a novel class of selective COX - 2 inhibitors. *J. Med. Chem.*, 1998, 41(24): 4800 ~ 4818
- [14] Li J J, Norton M B, Renhard E J, et al. Novel terphenyls as selective cyclooxygenase - 2 inhibitors and orally active antiinflammatory agents. *J.*

- Med. Chem., 1996, 39: 1846~1856
- [15] Marnett L J, Kalgutkar A S. Design of selective inhibitors of cyclooxygenase - 2 as nonulcerogenic anti - inflammatory agents. *Current Opinion in Chemical Biology*, 1998, 2: 482~490
 - [16] Gars K R, Gallraith W, Romsn R J, et al. Antiinflammatory and safety profile of DuP 697, a novel orally effective prostaglandin synthesis inhibitor. *J. Pharmacol. Exptl. Ther.*, 1990, 254(1): 180~187
 - [17] Magolda R L, Batt D, Covington M B, et al. Structure - activity relationships with a novel series of selective cyclooxygenase - 2 inhibitors. *Inflamm. Res.*, 1995, 44(suppl. 3): A274
 - [18] Reitz D B, Li J J, Norton M B, et al. Selective cyclooxygenase inhibitors: novel 1,2 - diarylcyclopentenones are potent and orally active COX - 2 inhibitors. *J. Med. Chem.*, 1994, 37(23): 3878~3881
 - [19] Penning T D, Talley J J, Bertenshaw S R, et al. Synthesis and biological evaluation of the 1,5 - diarylpyrazole class of cyclooxygenase - 2 inhibitor: identification of 4 - [5 - (4 - methylphenyl) - 3 - (trifluoromethyl) - 1 - yl] benzenesulfonamide (SC - 58635, Celecoxib). *J. Med. Chem.*, 1997, 40(9): 1347~1365
 - [20] Futaki N, Arai I, Hamasaka Y, et al. Selective inhibition of NS - 398 on prostanoid production in flamed tissue in rat carrageenan - air - pouch inflammation. *J. Pharm. Pharmacol.*, 1993, 45(8): 753~755
 - [21] Masferrer J L, Zweifel B S, Manning P T, et al. Selective inhibition of inducible cyclooxygenase 2 in vivo is antiinflammatory and nonulcerogenic. *Proc. Natl. Acad. Sci. U.S.A.*, 1994, 91: 3228~3232
 - [22] Huff R, Collins P, Kramer S. A structural feature of N - [2 - cyclohexyloxy - 4 - nitrophenyl] methanesulfonamide (NS - 398) that governs its selective and affinity for cyclooxygenase - 2 (COX - 2). *Inflamm. Res.*, 1995, 44(suppl. 2): 145~146
 - [23] Hirschelmann R, Hentschel M, Rickinger O. CGP 28238, a new potent nonsteroidal antiinflammatory agent: its relation to arachidonic acid metabolism. *Agent Actions*, 1991, 32(11): 54~55
 - [24] Bendele A M, Benslay D N, Hom J T, et al. Antiinflammatory activity of

- BF389, a Di - T - butylphenol, in animal models of arthritis. *J. Pharmacol. Exp. Ther.*, 1992, 260(1): 300 ~ 305
- [25] Katritzky A R, Maran U, Lobanov V S, Karelson M. Perspective: structurally diverse quantitative structure - property relationship correlations of technologically relevant physical properties. *J. Chem. Inf. Comput. Sci.*, 2000, 40(1): 1 ~ 18
- [26] Robinson D D, Winn P J, Lyne P D, Richards W G. Self - organizing molecular field analysis: a tool for structure - activity studies. *J. Med. Chem.*, 1999, 42: 573 ~ 583
- [27] Marot C, Chavatte P, Lesieur D. Comparative molecular field analysis of selective cyclooxygenase - 2 (COX - 2) inhibitors. *Quant. Struct. - Act. Rel.*, 2000, 19(2): 127 ~ 134
- [28] Singh P, Kumar R. 1,2 - Diarylimidazoles as inhibitors of cyclooxygenase - 2: A quantitative structure - activity relationship study. *J. Enzym. Inhib.*, 1999, 14(4): 277 ~ 288
- [29] Singh P, Kumar R. Novel inhibitors of cyclooxygenase - 2: The sulfones and sulfonamides of 1,2 - diaryl - 4,5 - difluorobenzene. Analysis of quantitative structure - activity relationship. *J. Enzym. Inhib.*, 1998, 13(6): 409 ~ 417
- [30] Kumar R, Singh P. Diarylspiro [2,4] heptenes as selective cyclooxygenase - 2 inhibitors: A quantitative structure - activity relationship analysis. *Indian J. Chem. B*, 1997, 36(12): 1164 ~ 1168
- [31] Wilkerson W W, Copeland R A, Covington M B, et al. Cyclooxygenase - 2 inhibitory 2 - substituted - 4,5 - diarylpyrroles. *Med. Chem. Res.*, 1995, 5(5): 399 ~ 408
- [32] Kalgutker A S, Crews B C, Rowlinson S W, et al. Biochemically based design of cyclooxygenase - 2 (COX - 2) inhibitors: Facile conversion of non-steroidal antiinflammatory drugs to potent and highly selective COX - 2 inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, 2000, 97(2): 925 ~ 930
- [33] Liu S S, Yin C S, Shi Y Y, Cai S X, Li Z L. MEDV - 13 for QSAR studies on the COX - 2 inhibition by indomethacin amides and esters. *Chin. J. Chem.*, 2001, 19(8): 751 ~ 756

- [34] Black W C, Brideau C, Chan C C, et al. 2,3 - Diarycyclopentenones as orally active, highly selective cyclooxygenase - 2 inhibitors. J. Med. Chem. , 1999, 42: 1274 ~ 1281
- [35] Song Y, Connor D T, Doubleday R, et al. Synthesis, structure - activity relationships, and in vivo evaluations of substituted di - tert - butylphenols as a novel class of potent, selective, and orally active cyclooxygenase - 2 inhibitors. 1. Thiazolone and oxazolone series. J. Med. Chem. , 1999, 42: 1151 ~ 1160
- [36] Puig C, Crespo M I, Godessart N, et al. Synthesis and biological evaluation of 3,4 - diaryloxazolones: a new class of orally active cyclooxygenase - 2 inhibitors. J. Med. Chem. , 2000, 43: 214 ~ 223

附录

发表与博士论文有关的学术论文

1 被 SCI 收录的学术论文

1. **Liu S S**, Liu H L, Xia Z N, Cao C Z, Li Z L. Molecular Distance - Edge Vector (μ): An Extension from Alkanes to Alcohols. J. Chem. Inf. Comput. Sci. 1999,39(6):951~957
2. **Liu S S**, Xia Z N, Liu Y, Cai S X, Li Z L. An atomic electronegative distance vector and carbon - 13 nuclear magnetic resonance chemical shifts of alcohols and alkanes. Chin. J. Chem. 2000,18(2):165~174
3. **Liu S S**, Liu H L, Xia Z N, Cao C Z, Li Z L. Molecular Distance - Edge Vector (μ) and Chromatographic Retention Index of Alkanes. J. Chin. Chem. Soc. 2000,47(3):455~460
4. **Liu S S**, Cai S X, Cao C Z, Li Z L. Molecular Electronegative Distance Vector (MEDV) Relating to 15 Properties of Alkanes. J. Chem. Inf. Comput. Sci. 2000,40(6):1337~1348
5. **Liu S S**, Yin C S, Cai S X, Li Z L. Structural Vector Description and Estimation of Normal Boiling Points for 66 Aromatic Hydrocarbons. Chin. Chem. Lett. 2000,11(12):1089~1092
6. 刘树深, 刘堰, 蔡绍哲, 李志良. 一个新的分子电性距离矢量 (MEDV). 化

学学报,2000,58(11):1353~1357

7. **Liu S S**, Yin C S, Cai S X, Li Z L. QSAR Study of Steroid Benchmark and Dipeptides Based on MEDV - 13. J. Chem. Inf. Comput. Sci. 2001, 41 (2):321~329
8. **Liu S S**, Liu H L, Yu B M, Cao C Z, Li Z L. Investigation on Quantitative Relationship between Chemical Shift of Carbon - 13 Nuclear Magnetic Resonance Spectra and Molecular Topological Structure Based on a Novel Atomic Distance - Edge Vector (ADEV). J. Chemom. 2001,15(5):427~438
9. **Liu S S**, Yin C S, Cai S X, Li Z L. A Novel MHDV Descriptor for Dipeptide QSAR Studies. J. Chin. Chem. Soc. 2001,48(2):253~260
10. **Liu S S**, Yin C S, Shi Y Y, Cai S X, Li Z L. MEDV - 13 for QSAR Studies on the COX - 2 Inhibition by Indomethacin Amides and Esters. Chin. J. Chem.,2001,19(8):751~756
11. **Liu S S**, Yin C S, Cai S X, Li Z L. Molecular structural vector description and retention index of polycyclic aromatic hydrocarbons. Chemom. Intell. Lab. Syst. 2002,61:3~15
12. **Liu S S**, Yin C S, Wang L S. Combined MEDV - GA - MLR Method for QSAR of Three Panels of Steroids, Dipeptides, and COX - 2 Inhibitors. J. Chem. Inf. Comput. Sci. 2002,42(3):749~756
13. **Liu S S**, Yin C S, Wang X D, Wang L S. QSAR Studies on Dipeptides Based on A Combinatorial MHDV - GA - MLR method. J. Chin. Chem. Soc. 2002,49(6):1089~1096
14. **Liu S S**, Yin C S, Wang L S. MEDV - 13 for QSRR of 62 Polychlorinated Naphthalenes. Chin. Chem. Lett. 2002,13(8):791~794
15. **Liu S S**, Liu H L, Yin C S, Wang L S. VSMP: A Novel Variable Selection and Modeling Method based on the Prediction. J. Chem. Inf. Comput. Sci. 2003,43(3):964~969
16. **Liu S S**, Cui S H, Yin D Q, Shi Y Y, Wang L S. QSAR Studies on the COX - 2 Inhibition by 3,4 - Diarylcyclohexanones based on MEDV Descriptor. Chin. J. Chem. 2003,21(11):1510~1516
17. Liu Y, **Liu S S**, Cui S H, Cai S X. A Novel Quantitative Structure - Biodegradability Relationship (QSBR) of Substituted Benzenes Based on MHDV

Descriptor. J. Chin. Chem. Soc. 2003,50(2):319~324

18. **Liu S S**, Cui S H, Wang L S. Prediction of the Aqueous Solubilities of Polychlorinated Biphenyls. Chin. Chem. Lett. 2004,15(4):467~470

2 其他论文

1. 刘树深, 夏之宁, 余般梅, 李志良. 无环醇原子电距矢量及核磁共振碳谱化学位移模拟. 波谱学杂志, 1999,16(5):429~440
2. **Liu S S**, Liu H L, Shi Y Y, Wang L S. QSAR of Cyclooxygenase - 2 (COX - 2) Inhibition by 2,3 - Diarylcyclopentenones Based on MEDV - 13. Internet Electron. J. Mol. Des. 2002,1(6):310~318
3. **Liu S S**, Cui S H, Shi Y Y, Wang L S. Novel Variable Selection and Modeling Method based on the Prediction for QSAR of Cyclooxygenase - 2 Inhibition by Thiazolone and Oxazolone Series. Internet Electron. J. Mol. Des. 2002,1(11):610~619