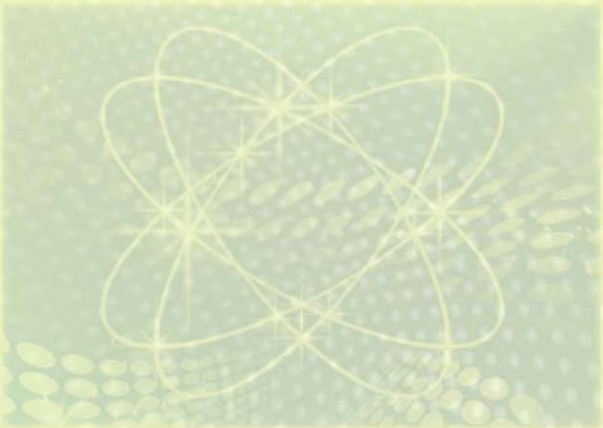
量子化学与人工智能计算在分子键能中的应用



序

从二十世纪六十年代后期开始计算机技术高速发展，不仅带动了各门学科的发展速度，同时也催生了一批新的学科，化学信息学就是其中之一，即利用信息学方法来解决化学问题。利用计算机技术来分析问题，解决问题，或采集、分类、处理化学研究中的产生的大量的数据，提取其中包含的重要信息。

理论计算化学建立在物理、数学理论方法基础上，建立分子的结构和物理化学性质的方程，从而求解分子各种性质。但是由于分子越大分子结构也是复杂，求解方程难度大且花费时间多，所以有时只能牺牲计算精度来达到实现实际计算的目的。近二十年来，随着计算机技术水平提高，理论计算化学的发展迅速，但是从提升理论方程的角度来提高理论计算的精度，还是效率很低，对于方法的使用还是有很大的局限性。

机器学习方法能够通过学习掌握数据的特点，建立输入和目标值之间的关系，然后利用这个关系就能够对新的输入进行预测得到新的目标值。这种方式可以绕开求解输入和输出之间方程的细枝末节，一样可以得到求解方程得到结果，而且利用此方法比

在求解函数方程要简单、高效。

本书的研究是利用机器学习方法提高量子化学计算结果的精度。第一章介绍了有关分子键能的基本知识。第二章介绍了量子化学计算方法的基本理论和常用基组的构成。第三章、第四章、第五章和第六章分别研究了利用基于平均影响值的反向传播神经网络、基于灰色关联分析和主成分分析的广义回归神经网络、基于自组织特征映射神经网络的径向基神经网络以及基于层次聚类和蚁群聚类优化的极限学习机方法来提高密度泛函方法在不同基组水平下计算出来的键均裂能的精度。研究展示了机器学习方法对于提高量子化学计算精度的有效性，一个实现高精度计算高效、快捷的方法。印证了结合多学科方法可以为提高计算效率和精度的提供新的途径。

胡丽红

2013 年 4 月 10 日于东北师范大学净月校区

目　录

第一章 绪 论

一、研究背景与研究意义

（一）研究背景

量子力学是20世纪最重要的科学发现之一。在量子力学基础上发展起来的理论物理，量子化学及相关的计算，为我们开辟了通向微观世界的又一个途径[1]。量子化学计算的一大优势在于它可以先于实验来预测物质的性质或实验上至今无法测得的一些物理量及无法观测到的反应过程。量子化学是研究分子微观结构、性质和分子间相互作用的最基础学科[2-3]。

量子化学方法直接讨论分子结构与性能二者之间的关系，所得参数本身含有分子的几何结构，化学结构和电子结构等信息。从理论上讲，通过量子化学理论能够对分子的电子结构和几何结构进行计算，得到分子的各种化学参数。计算得到的参数物理意义明确，因而可以从理论上对分子的性质进行解释和预测，已成为许多涉及研究分子层次相关学科的基础。相比传统的经验参

数，量子化学参数对化合物结构的描述更加全面、细致、准确，物理意义更加明晰，理论性更强，因此可以根据量子化学的计算结果进一步揭示物质的结构与性质之间的关系。

物质的化学结构与性质的问题是我们学习和研究物理、化学、材料、能源、环境和太空等自然学科过程中的重要问题。物质是由原子构成。原子或原子团之间的相互作用形成分子，这种相互作用力就是化学键。化学键可分为共价键、离子键、配位键、金属键、氢键、静电相互作用、范德华力和表面键等。通过化学键可以形成各种各样的分子、“超分子”、络合物或簇合物。分子中化学键的强弱可以用来度量分子化学稳定性的高低。多原子分子中含有多个化学键，每个化学键的键能不一定相同，有强弱之分。化学键能总和一定是正值，表示化学键断裂时，需要外界提供能量，键能值越高，表示吸收越多的能量，才能断裂，反之，键能值越低，表示断裂该键所需要的能量就越少。

近年来密度泛函理论[4-7]的迅速发展，为复杂体系的研究提供了广阔天地。然而并不是所有的计算结果都是十分精确地，特别是对于复杂分子或者较大的系统[8]。导致这种局限性的主要原因是该计算方法本身采用固有的近似引起的。密度泛函理论计算方法的精度主要由所使用的交换相关函数决定的[9]，而它的准确的形式仍然是未知的，所有这些就导致了理论计算的误差。现在广泛应用的高精度方法采用大基组就会局限于计算小分子，对于中到大分子的计算，经常采用低精度的理论方法和小基组，虽然

节省了机时，但由于在计算中有电子相关效应和各种内在的近似，往往会导致较大的系统误差，有的误差已经超过了化学计算精度的极限值。所以人们开始寻求其他科学合理的计算方法来提高计算精度。

量子化学方法已经超过了仅仅验证实验值的水平，它能够在实验值不知道或不确定的情况下来预测化学键能，然而并不是所有的计算结果都十分精确，导致这种局限性的主要原因是计算方法本身固有的近似。因此，近 10 年来，很多统计学方法被用来提高量子化学的计算精度。先由量子化学方法计算得到相关的分子属性，然后应用统计方法建立实验值与计算值之间的关系。这些统计改进方法主要包括线性方法，如线性回归[10-13]等，和非线性方法如神经网络[14-36]等。尽管多元线性回归方法简单直观，但是对于使用相同的物理参数，神经网络可以较好的解决复杂的难以用数学公式建模的非线性问题[33]。最近，支持向量机是由 Vapnik 首先提出来的[37]，它具有很高的泛化能力，避免局部极小，自动获得网络的拓扑结构和较低的工作量。因此，它已被广泛关注和应用，如模式识别，图形处理，基因表达和选择，时间系列预测等。在化学信息学中，SVM 主要应用在多元校正[38-39]和分子的物理化学性质预测[40-57]。

（二）研究意义

研究利用多种人工智能方法或机器学习方法来校正量子化学

的理论计算键能值，以提高分子键能的量子化学计算精度。这种利用人工智能方法来迂回解决量子化学计算精度问题，避免了使用高级的量子化学方法和超大基组花费大量的时间，就能得到的高精度的计算结果。所以预期能在较少的机时和计算资源下得到比较精确的计算结果，或者在现有计算条件下，预测目前计算能力达不到的精确结果，并有效的指导实验。这种人工智能与量子化学计算结合的组合型计算方法能够互取所长，实现在简单的物理参数下，减小理论计算中所固有的近似所带来的系统误差，为准确、快捷地预测分子性质提供了一种新的研究手段。人工智能方法在量子化学数据分析中的处理应用，也将有助于发现新的规律和创造新的物质结构。

化学键能是分子的热化学性质之一，它可以用来度量分子的化学稳定性的高低和控制了许多反应速度的快慢并确定了相应化学反应的机理，但是化学键能的测量却是很困难的，实验过程中的微小误差都可能使键能值的测定产生很大的误差。因此精确地预测分子键能是计算化学领域的一个重要问题。

量子化学方法是建立在量子力学基础理论上，来研究分子微观结构、性质和分子间相互作用等理论计算方法。大量的计算和实验的结果比对说明，量子化学计算方法真正抓住了分子的物理本质，所以量子化学方法在近二十年的的发展非常迅速，作为与实验互补方法或代替某些实验研究得到广泛应用。

量子化学计算的一大优势在于它可以先于实验来预测物质的

性质或实验上至今无法测得的一些物理量及无法观测到的反应过程。量子化学计算方法通过计算分子体系的电子波函数方程，可以求得偶极矩、极化率、热力学、激发态等分子性质，但是由于计算条件的局限，不得不在求解波函数方程过程中引进一些近似，以实现方程的求解。随着化学和相关学科的发展，研究手段的不断提高，研究的体系也逐渐从小分子和简单体系转向大分子和实际的聚集或凝聚态体系，随之而来的是如何解决计算精度的问题或是发展更适合研究复杂体系的方法。对于结构规整的较大体系，量子化学方法也能给出合理的描述，但是对于计算复杂的体系，现有的方法还是很难进行或者计算精度太差，远远超过了进行科学研究所要求的误差范围。量子化学方法的计算误差主要来自于电子相关效应和计算采用的基组，因为实际计算中所选择的基组有限，不能包含分子全部的空间轨道，所以给理论计算带来了误差。现在广泛应用的高精度方法采用大基组就会局限于计算小分子，对于中到大分子的计算，经常采用低精度的理论方法和小基组，虽然节省了机时，但由于在计算中有电子相关效应和各种内在的近似，往往会导致较大的系统误差，使计算结果失去其应用价值，所以人们一直在寻求合理的计算方法来提高计算精度。改进量子化学计算方法提高量子化学的计算精度是理论计算的重要目标之一。通过改进量子化学计算方法来提高精度，难度很大，修改方程，减少近似，应用先进完备的算法等手段都给计算条件带来更高的要求，即使结果的精度提高很小，都非常困

难，更不用说得到高精度结果了。

人工智能方法模仿人脑结构或思维机理，通过计算机实现智能的方法，目前得到了愈加广泛的应用，它在机器人控制，经济政治决策，控制系统，仿真系统中，医学诊断，股票交易，法律，科学发现等领域都得到应用或利用。在本研究当中人工智能方法将用于校正量子化学计算中的误差，通过人工智能方法建立计算值和实验值之间的关系，绕过严格求解波函数方程的难题，从而得到高精度的计算结果。

虽然理论计算值和实验值之间没有严格的、确定性的函数关系，但可以找出最能代表它们之间关系的表达形式。量子化学计算的误差不可避免，但是量子化学计算可以捕捉到分子的重要物理本质属性。理论计算和实验结果之间的关系主要取决于主导属性的特性，并在一定程度上和分子的其他属性也有关系。理论计算和实验结果之间的误差在很大程度上都来源于第一原理方法的近似，尽管通过第一原理很难确定理论计算和实验结果之间的关系，但是这种数量关系可以轻易通过人工智能方法得出。事实上，一种现象常常是与多个因素相联系的，由多个物理参数的最优组合共同来预测或估计实验值，比只用一个或随机搭配物理参数进行预测或估计更有效，更符合实际。

以对一氧化氮（NO）键能的研究为例，研究中根据所选数据集来确定物理参数，先应用灰色关联分析（GRA）来筛选物理参数，再应用主成分分析（PCA）来优化所选择的物理参数。可

见基于 GRA 和 PCA 的广义回归神经网络（GRNN）（GP-GRNN）有更好的适应能力和校正效果。即量子化学计算方法 B3LYP/6-31G（d）计算后的均方根误差为 5.31kcal mol^{-1}，全参数下 GRNN（F-GRNN）校正后的均方根误差为 0.49kcal mol^{-1}，当物理参数只经过 GRA 处理后（G-GRNN）以及在这基础上继续进行 PCA 处理后（GP-GRNN）通过 GRNN 校正后的均方根误差分别减小到 0.39kcal mol^{-1} 和 0.31kcal mol^{-1}。由于优化所选择的物理参数，避免可能出现的信息存在重叠现象，以及可能出现运算不稳定和病态矩阵等问题，即权重分配更加合理，避免信息的冗余，去掉了噪声，提高了数据的质量，使计算结果的精确度得到了进一步的提高。

综上所述，由于量子化学计算的误差较大或者计算量巨大，单一的量子化学方法难以得到满意的结果。将人工智能方法应用于量子化学计算中研究分子的键能，在国际上也尚未有人深入研究。尤其是针对复杂的大分子体系，更没有一种通用的有效的方法。本研究针对具体的研究对象选择合理的物理参数，对所选参数的重要性给出合理的解释，并深入探讨各种人工智能方法在具体研究对象中的优缺点，以及影响量子化学计算精度因素的分析。本研究着眼于效率的同时，更兼顾其通用性，在理论研究和应用研究方面都具有重要的意义。

二、研究目标与内容

（一）研究目标

主要研究利用多种人工智能或机器学习方法来校正量子化学理论计算的键能值，以提高分子键能的量子化学计算精度，这样不仅可以解决单纯利用实验或量子化学理论计算得到高精度的键能值的高昂成本问题，而且用此建立起来的结构和性能的高精度函数关系预测新型高性能材料分子结构。

（二）研究内容

根据热化学和热力学的习惯，化学键能被定义为键离解反应的焓变，它是分子的热化学性质之一。通常把键离解能（焓）称为键能。键能值越高，表示分子的化学稳定性越高；键能值越低，表示分子的化学稳定性越差。在化学热力学里，化学键能值的高低，基本上控制了许多反应速度的快慢，也确定了相应化学反应的机理。精确测量不同分子、离子、自由基、超分子、络合物、氢键合物以及表面键合物中化学键能是非常困难的工作。尽管测量键能的实验方案和手段很先进，但结果却不尽如人意。其原因是化学键离解以后的碎片大多数是寿命极短、浓度很低的反

应活性中间体，如自由基等。要捕捉和检测这些活性碎片很困难，如要定量测定这些痕量碎片的浓度并研究相应键离解过程与温度的定量关系就更难了。

应用多种人工智能算法预测分子键能值，探讨各种人工智能算法的优越性，并寻找选择主要的分子物理参数。期望能在较少的机时和计算资源下得到更加精确的计算结果，或者在现有计算条件下，预测目前计算能力达不到的计算精度。具体研究内容主要包括以下几个部分：

1）选择数据集

建立大量分子键能的实验数据，构建分子模型，应用量子化学计算方法进行理论计算，从输出结果中获取充足的特征参数。现在已经建立了 92 个数据分子模型，分子体系的理论计算值已经求得。

2）选择不同精度的量子化学计算方法

根据具体数据集的需要选择不同的量子化学计算方法进行理论计算，并着重研究各量子化学计算方法对不同类型数据的选择，在同一量子计算方法下研究不同的基组对于不同类型数据的选择。选用不同的量子化学计算方法来寻找基于智能算法的最优组合，拟选用的量化方法有：密度泛函理论等。还将选用不同的基组来寻找基于人工智能算法的最优组合，拟选用的基组有：6-31G（d），STO-3G 等。