XXVIII Congresso {virtual} de Iniciação Científica da Unicamp 2020

AVALIAÇÃO DE ESQUEMAS DE OTIMIZAÇÃO PARA O DESENVOLVIMENTO DE MÉTODO COMPOSTO EM CÁLCULOS DE ENTALPIAS DE FORMAÇÃO

Pedro Giraldi Faccin Prof. Dr. Rogério Custodio

1 Objetivos

Desenvolver um método composto baseado na teoria G3 no que se refere ao tratamento da correlação eletrônica, funções de base e efeitos térmicos com exatidão em relação entalpias de formação experimentais acuradas. A otimização dos parâmetros será realizada com método tradicional Simplex e, posteriormente, baseado em redes neurais artificiais. O método proposto terá um conjunto de treinamento e um conjunto de validação externa para o teste dos parâmetros otimizados, aumentando assim a confiabilidade da margem de erro alcançada. Serão desenvolvidos scripts e programas em Python para a manipulação de dados, otimização de processos e implementação dos métodos de cálculo a serem utilizados.

2 Introdução

2.1 Métodos

O cálculo de propriedades eletrônicas, termoquímicas e espectroscópicas de átomos e moléculas constitui um grande desafio dentro da química quântica para o desenvolvimento de métodos que permitam a obtenção de valores compatíveis com dados experimentais acurados. Na literatura existem diversos métodos para a obtenção dessas propriedades para substâncias de dimensões variáveis utilizando cálculos ab initio, modelos baseados na teoria do funcional de densidade e métodos compostos. A elaboração e otimização de tais métodos torna possível à obtenção de resultados que se assemelhem aos dados experimentais com custo computacional cada vez menor. Os métodos compostos representam uma via para a obtenção desses resultados utilizando a combinação de cálculos ab initio de nível inferior com resultados compatíveis ao de cálculos únicos de alto nível que demandam grande esforço computacional.

Os mais bem-sucedidos e amplamente utilizados métodos compostos são os denominados Gaussian-n ou simplesmente Gn desenvolvidos por Pople,Curtiss e col. [1–5] . A teoria Gaussian-3 ou G3 é a terceira na série de métodos Gn [4] e a energia final é definida por:

$$E_{G3} = E_{MP4/6\text{-}31G(d)} + \Delta E_{(+)} + \Delta E_{(2df,p)} + \Delta E_{QCI} + \Delta E_{G3\,large} + E_{SO} + E_{ZPE} + E_{HLC}$$

Em que a energia de referência $E_{MP4/6\text{-}31G(d)}$ é modificada pelas seguintes correções:
(a) $\Delta E_{(+)} = E_{MP4/6\text{-}31+G(d)} - E_{MP4/6\text{-}31G(d)}$ para funções difusas,
(b) $\Delta E_{2df,p} = E_{MP4/6\text{-}31G(2df,p)} - E_{MP4/6\text{-}31G(d)}$ para funções de polarização em átomos com exceção do hidrogênio e funções-p em hidrogênios,
(c) $\Delta E_{QCI} = E_{QCISD(T)/6\text{-}31G(d)} - E_{MP4/6\text{-}31G(d)}$ para efeitos de correlação eletrônica,
(d) $\Delta E_{G3\ large} = E_{MP2(full)/G3\ large} - E_{MP2/6\text{-}31G(2df,p)} - E_{MP2/6\text{-}31+G(d)} + E_{MP2/6\text{-}31G(d)}$ para funções de bases extensas,
(e) E_{SO} para correção de spin-órbita extraída de experimentos atômicos ou cálculos teóricos
(f) para levar em conta outros efeitos de correlação de alto nível $E_{HLC} = -An_\beta - B(n_\alpha - n_\beta)$ para moléculas e $E_{HLC} = -Cn_\beta - D(n_\alpha - n_\beta)$ para átomos, em que n_α e n_β são o número de elétrons de valência com spins alfa e beta, respectivamente, e A, B, C e D são parâmetros otimizados para a produzir o menor erro absoluto médio possível em relação a valores experimentais [6].

Nos métodos Gn chama a atenção que a correção E_{HLC} tem por objetivo corrigir outras deficiências no cálculo dos termos de correlação e função de base. Isto sugere que talvez seja possível substituir o ajuste E_{HLC} por fatores de escalamento realizados diretamente nos termos de correção presentes, por exemplo, na energia G3. Em outras palavras, a expressão para a energia G3 final seria dada por:

$$E_{G3} = E_{MP4/6\text{-}31G(d)} + a\Delta E_{(+)} + b\Delta E_{(2df,p)} + c\Delta E_{QCI} + d\Delta E_{G3large} + eE_{SO} + fE_{ZPE}$$

Uma vantagem em utilizar esta formulação é que podemos ajustar os termos de correlação eletrônica e função de base diretamente e constatar através dos valores otimizados de cada parâmetro a importância do efeito de cada termo no cálculo da propriedade de interesse, como a entalpia de formação.

A proposta de se substituir a correção HLC por termos de escalamento aplicados diretamente em correções de correlação e funções de base não é nova. De fato, o método G3S, uma versão modificada do método G3, introduziu parcialmente esta ideia seguindo a ótica de trabalhos semelhantes realizados por Gordon e Truhlar [7,8] no escalamento de todos os termos de correlação (SAC), Seigbahn e col. no método de correlação parametrizada PCI-X [9] e no método de multi-coeficiente de correlação de Truhlar e col. (MCCM) [10,11]. No método G3S foram avaliados ajustes considerando diferentes efeitos de correlação e ajustes de funções de base que compõe o método G3. Com um número máximo de 6parâmetros o método G3S alcançou nível de exatidão experimental de 0.99 kJ mol^{-1} no cálculo de diferentes tipos de energias para moléculas do conjunto G2/97 que contém 299 energias experimentais.

Curiosamente, no método G3S não há um conjunto de validação externa,isto é, todo o conjunto G2/97 foi utilizado para a otimização dos parâmetros, os valores finais desses parâmetros estão sujeitos a overfitting.

2.2 Redes neurais

No seguimento de criação de modelos de machine learning destinados a previsão de propriedades eletrônicas, um dos algoritmos que têm ganhado mais popularidade nos últimos anos é a Rede Neural Artificial [12–16]. Redes neurais artificiais são modelos computacionais inspirados no sistema nervoso de seres vivos [17]. Assim, essas redes podem ser definidas como um conjunto de unidades de processamento, caracterizadas por neurônios artificiais, que são interligados por diversas interconexões (sinapses artificiais), sendo representadas por vetores/matrizes de pesos sinápticos. Essas interconexões são prolongadas e reprogramadas através de camadas dentro da estrutura da rede.

Recentemente, redes neurais artificiais têm se tornado bastante populares e úteis de forma geral na criação de modelos de classificação, agrupamento,reconhecimento de padrões e previsões em muitos campos da

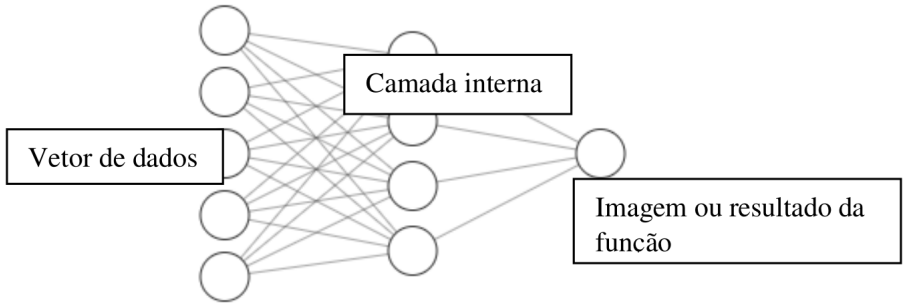

Figura 1: Modelo de uma rede neural do tipo feed forward

ciência,ganhando cada vez mais espaço competitivo entre os demais algoritmos em estado da arte [18] . As propriedades aproximativas das redes neurais do tipo feed forward já foram amplamente estudadas e comprovadas serem bastante gerais caracterizando esse algoritmo como um aproximador universal. Uma rede clássica, contendo duas camadas, por exemplo, pode aproximar qualquer função contínua dentro de um domínio relativamente compacto com qualquer valor arbitrário de exatidão dado que o framework da rede possua unidades internas ou neurônios o suficiente. O problema central das redes neurais artificiais é identificar quais são os melhores parâmetros a serem empregados na criação de um dado modelo à vista das características específicas do problema e do conjunto de treinamento disponível.

3 Métodos computacionais

3.1 Entalpias de formação

O procedimento padrão para o cálculo das entalpias de formação pode ser expresso através de uma sequência de equações bem conhecidas na literatura. A sequência de equações apresentada por Curtiss e col. [19] inicia-se através da energia de atomização:

$$D_0 = \sum xE_{\text{átomos}} - E_{\text{molécula}} - E_{ZPE}$$

Sendo $\sum xE_{\text{átomos}}$, o somatório das energias de cada átomo presente na molécula, e o termo $E_{\text{molécula}}$ a energia da molécula e E_{zpe} a energia de ponto-zero. Tanto o termo $E_{\text{átomos}}$ quanto o termo $E_{\text{moléculas}}$ serão obtidos através das equações definidas pelo método G3 modificado.

Com base na energia de atomização, pode-se calcular a entalpia de formação na temperatura de 0 K:

$$\Delta_f H^0(0\text{ K, molécula}) = \sum x\Delta_f H^0(0\text{ K, átomos}) - D_0$$

O primeiro termo à direita da igualdade, $\Delta_f H^0(0\text{ K, molécula})$, simboliza o somatório das entalpias de formação a 0 K dos elementos químicos. Os valores para esses termo são obtidos através de dados experimentais.

Finalmente, a entalpia de formação a 298 K é calculada através de correções térmicas:

$$\Delta_f H^0(0\text{ K, molécula}) = \Delta f H^0(0\text{ K, molécula}) + [H^0(298\ K) - H^0(0\ K)]_{\text{molécula}} - \sum x[H^0(298\ K) - H^0(0\ K)]_{\text{átomos}}$$

O termo $[H^0(298\ K) - H^0(0\ K)]_{\text{átomos}}$ é a correção térmica dos elementos. Já $[H^0(298\ K) - H^0(0\ K)]_{\text{molécula}}$ é obtido a partir da energia de ponto-zero e do termo H_{corr}, denominado correção térmica da entalpia, também fornecido pelos cálculos de frequência:

$$[H^0(298\ K) - H^0(0\ K)]_{\text{molécula}} = H_{corr} - E_{ZPE}$$

A determinação de H_{corr} envolve contribuições de movimento das energias translacionais, vibracionais, rotacionais e eletrônica, além da constante de Boltzmann e da temperatura, sendo expressa como:

$$H_{corr} = E_{eletr} + E_{vib} + E_{rot} + E_{trans} + k_b T$$

3.2 Otimização dos parâmetros

Aa primeira abordagem no processo de ajuste dos parâmetros da equação E_{G3} parametrizada será utilizar diretamente o método Simplex modificado de Nelder e Mead [21] para minimizar o erro médio absoluto das entalpias calculadas em relação aos dados experimentais.

Na segunda abordagem, o uso das redes neurais artificiais para realizar o agrupamento de moléculas de acordo com a similaridade destas no que se refere ao tratamento ótimo da correlação eletrônica e dos efeitos de aumento de base, podemos considerar a otimização dos parâmetros relacionados com as diferentes correções para cada molécula para permitir o agrupamento das moléculas cujos parâmetros sejam semelhantes. Para tanto, os parâmetros ótimos para cada uma das moléculas selecionadas serão obtidos e registrados em uma matriz **P** convergindo a 0 o erro associado à entalpia deformação calculado de cada molécula em relação ao valor experimental. A otimização dos parâmetros inerentes da rede será realizada utilizando o algoritmo baseado no método simplex modificado de Nelder e Mead [21].

Como pré-tratamento, os dados da matriz de parâmetros **P** serão processados centrando cada coluna de **P** na média. O agrupamento das moléculas será realizado com diferentes números de etapas por meio de uma análise de agrupamentos hierárquicos (HCA) realizada na matriz **P**. Em cada etapa serão construídos **N** grupos correspondentes, isto é, 1 grupo na primeira etapa (onde não há agrupamento de fato), 2 grupos na segunda etapa e assim sucessivamente. O método Ward de agrupamento será empregado de forma a minimizar a variância dentro de cada grupo. Assim, o conjunto de moléculas de referência será subdividido em **N** grupos para posterior otimização local dos parâmetros e criação do modelo de classificação.

Finalmente, em cada etapa E_n de agrupamento, para cada grupo **N** será retirado em torno de 20% das moléculas para a criação de um conjunto de teste do modelo de classificação criado. Para constituir esta parcela de 20% de cada grupo serão selecionadas moléculas de forma totalmente aleatória e nenhuma dessas moléculas deve ter participado de qualquer das etapas de modelagem ou otimização do método composto desenhado acima, sendo assim completamente estranhas ao modelo e método criados. Tendo criado para cada etapa E_n nos grupos pertencentes ao conjunto de treinamento e quais moléculas os compõem, a tarefa seguinte será criar uma matriz de dados **X** que contenha informações químicas através das quais seja possível treinar um

modelo de classificação. Visto que os grupos são definidos de acordo com a similaridade dos parâmetros que ajustam os termos de correlação eletrônica e função de base,a matriz de dados **X** será construída a partir das energias que constituem esses termos mais a contagem de átomos dos principais elementos representativos que constituem as moléculas do conjunto G3/05.

4 Resultados preliminares

A primeira fase do projeto é focada na implementação dos códigos e algorítimos citados. O cálculo da entalpia de formação das moléculas presentes no conjunto G3/05 já podem ser feitas de forma automatizada, já que os scripts para submissão, envio, análise e cálculo dos resultados estão prontos. Com esses programas prontos, foi possível atingir os mesmos resultados obtidos por Curtiss et al. [19] no quesito de error relativo nas entalpias de formação. Os algorítimos referentes a rede neural estão em fase de implementação, porém espera-se atingir erros médio absolutos menores dos até então encontrados.

5 Referências

1. Pople, J. A., Head-Gordon, M., Fox, D. J., Raghavachari, K. & Curtiss, L.A. Gaussian-1 theory: A general procedure for prediction of molecularenergies. J. Chem. Phys. 90, 5622–5629 (1989).
2. Shao, Y. et al. Advances in methods and algorithms in a modern quantumchemistry program package. Phys. Chem. Chem. Phys. 8, 3172–3191 (2006).
3. Blaudeau, J.-P., McGrath, M. P., Curtiss, L. A. & Radom, L. Extension ofGaussian-2 (G2) theory to molecules containing third-row atoms K andCa. J. Chem. Phys. 107, 5016–5021 (1997).
4. Curtiss, L. A., Raghavachari, K., Redfern, P. C., Rassolov, V. & Pople, J.A. Gaussian-3 (G3) theory for molecules containing first and second-rowatoms. J. Chem. Phys. 109, 7764–7776 (1998).
5. Curtiss, L. A., Redfern, P. C. & Raghavachari, K. Gn theory. WileyInterdiscip. Rev. Comput. Mol. Sci. 1, 810–825 (2011).
6. Curtiss, L. A., Raghavachari, K. & Pople, J. A. Gaussian-2 theory usingreduced Moller-Plesset orders. J. Chem. Phys. 98, 1293–1298 (1992).
7. Gordon, M. S. & Truhlar, D. G. Scaling all correlation energy inperturbation theory calculations of bond energies and barrier heights. J.Am. Chem. Soc. 108, 5412–5419 (1986).
8. Gordon, M. S. & Truhlar, D. G. Correlation balance in basis sets forelectronic structure calculations. Int. J. Quantum Chem. 31, 81–90 (1987).
9. Siegbahn, P. E. M., Blomberg, M. R. A. & Svensson, M. PCI-X, aparametrized correlation method containing a single adjustable parameterX. Chem. Phys. Lett. 223, 35–45 (1994).
10. Fast, P. L., Corchado, J. C., Sánchez, M. L. & Truhlar, D. G. Multi-Coefficient Correlation Method for Quantum Chemistry. J. Phys. Chem. A103, 5129–5136 (1999).
11. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Big DataMeets Quantum Chemistry Approximations: The Δ-Machine LearningApproach. J. Chem. Theory Comput. 11, 2087–2096 (2015).
12. von Lilienfeld, O. A., Ramakrishnan, R., Rupp, M. & Knoll, A. Fourierseries of atomic radial distribution functions: A molecular fingerprint formachine learning models of quantum chemical properties. Int. J. QuantumChem. 115, 1084–1093 (2015).
13. Montavon, G. et al. Machine learning of molecular electronic properties inchemical compound space. New J. Phys. 15, 095003 (2013).
14. Hansen, K. et al. Assessment and Validation of Machine LearningMethods for Predicting Molecular Atomization Energies. J. Chem. TheoryComput. 9, 3404–3419 (2013).
15. Schütt, O. & VandeVondele, J. Machine Learning Adaptive Basis Sets forEfficient Large Scale Density Functional Theory Simulation. J. Chem.Theory Comput. 14, 4168–4175 (2018).
16. da Silva, I. N., Spatti, D. H. & Flauzino, R. A. Redes Neurais ArtificiaisPara Engenharia e Ciências Aplicadas. (Artliber, 2016).
17. Abiodun, O. I. et al. State-of-the-art in artificial neural networkapplications: A survey. Heliyon 4, e00938 (2018).
18. Curtiss, L. A., Raghavachari, K., Redfern, P. C. & Pople, J. A.Assessment of Gaussian-2 and density functional theories for thecomputati
19. Narayanan, B., Redfern, P. C., Assary, R. S. & Curtiss, L. A. Accuratequantum chemical energies for 133 000 organic molecules. Chem. Sci.10, 7449–7455 (2019).
20. Nelder, J. A. & Mead, R. A Simplex Method for Function Minimization.Comput. J. 7, 308–313 (1965).