结合机器学习算法提高从头算方法对 HF/HBr/$H^{35}Cl$/$Na^{35}Cl$ 振动能谱的预测性能*

杨章章[1)] 刘丽[1)] 万致涛[1)] 付佳[1)†] 樊群超[1)‡]
谢锋[2)] 张燚[3)] 马杰[4)]

1) (西华大学理学院, 高性能科学计算四川省高校重点实验室, 成都 610039)
2) (清华大学核能与新能源技术研究院、先进核能技术协同创新中心、先进反应堆工程与安全教育部重点实验室, 北京 100084)
3) (国防科技大学, 前沿交叉学科学院, 长沙 410073)
4) (山西大学物理与电子工程学院, 量子光学与量子光学器件国家重点实验室, 太原 030006)

(2022 年 10 月 13 日收到; 2022 年 12 月 29 日收到修改稿)

高精度振动能谱蕴含着分子体系的大量量子特征, 是人们认识和操控分子的重要基础数据. 目前, 从头算方法在计算分子的振动能谱方面取得了大量成果, 但是仍然面临着精度和计算量上的挑战. 本文提出了一种综合从头算方法与机器学习算法进行能谱预测的新方法, 在提高振动能级精度的同时大幅降低了计算量. 针对 HF、HBr、$H^{35}Cl$ 和 $Na^{35}Cl$ 等卤素分子的研究结果表明, 相较于单独的 CCSD(T)/cc-pV5Z 计算方法, 新方法将误差减少了 50% 以上, 同时将计算量降低了一个数量级.

关键词: 振动能谱, 从头算, 机器学习, 卤素分子
PACS: 31.15.A–, 33.15.Mt, 33.20.–t, 33.20.Vq **DOI:** 10.7498/aps.72.20221953

引 言

卤化物在大气化学、控制臭氧丰度和钢材腐蚀等研究中有着重要作用, 一直备受关注[1–5]. 分子光谱蕴含着大量量子层面的分子信息, 其高精度数据对认识卤化物的精细结构有着重要的意义. 1971 年, Smart 和 Sheppard[6] 研究了卤化氢 (HCl, HBr 和 HI) 吸附在金属卤化物上的红外光谱. 1991 年, Blass 等[7] 测量了温度低于 83 K 时, HBr 和 HCl 在 LiF 单晶表面的红外光谱. 1997 年, Giorgi 等[8] 采用 High-n Rydberg-atom time-of-flight (HRTOF) 光谱技术记录了 HX (X = Cl, Br 和 I) 光解后的氢原子平移能量分布. 2006 年, Carvalho 等[9] 得到了 HBr 和 HCl 的振动能量分布. 2012 年, Sun[10] 通过实验记录了金属卤化物溶液的拉曼光谱. 诸如此类, 对卤化物光谱特性、振转能谱及其他物理化学性质的研究还有许多[11–13]. 在理论上, 常用从头算方法获取分子的势能, 并基于此进一步获取光谱数据. 在众多的从头算方法中, 耦合簇 (coupled-clusters, CC) 理论[14,15] 事实上已经成为了通用的能量计算手段. 然而, 由于耦合簇方法在处理量子多体效应时采用了多激发的方式, 这会带来沉重的计算负担. 实际可行的计算往往止步于全三激发 CCSDT 及其近似版本 CCSD(T)[16]. 因此, 很难从物理模型层面进一步提高 CC 方法的预测精度.

* 国家自然科学基金 (批准号: 11904295)、四川省科学技术计划 (批准号: 2021ZYD0050) 和极端光学协同创新中心开放研究基金项目 (批准号: KF2020003) 资助的课题.
† 通信作者. E-mail: fujiayouxiang@126.com
‡ 通信作者. E-mail: fanqunchao@sina.com

目前, 随着机器学习的发展, 其构建高维复杂函数的能力为物理学研究提供了诸多助力. 比如辅助光学表面杂质检测[17]、处理高分辨电镜图像[18]以及预测磁性材料基态磁矩[19]等. 本文将机器学习的高维函数建模能力与从头算方法结合, 获得了HF, HBr, $H^{35}Cl$ 和 $Na^{35}Cl$ 等卤化物的振动能级分布. 所得结果在精度和资源消耗上均优于 CCSD(T)/cc-pV5Z 方法.

1 理论与方法

1.1 从头算方法

对于双原子分子体系, 当给定电子位置r和原子核位置R时, 不显含时间影响的薛定谔方程可以记作:

$$\begin{aligned}&\hat{H}\psi\left(r_1,r_2,\cdots,r_n,R_1,R_2\right)\\&=E\psi\left(r_1,r_2,\cdots,r_n,R_1,R_2\right),\end{aligned}\tag{1}$$

其中 $\hat{H}$, E 和 $\psi\left(r_1,r_2,\cdots,r_n,R_1,R_2\right)$分别是哈密顿算符, 总能量和波函数.

按照波恩-奥本海默近似 (Born - Oppenheimer approximation), 可将电子和原子核的运动分开处理. 其中原子核在均匀势场中运动, 该场由原子核与处于稳定状态的电子相互作用产生. 结合双原子分子体系的几何对称性, (1) 式可以简化如下:

$$\begin{aligned}&\left\{-\frac{\hbar^2\mathrm{d}^2}{2\mu\mathrm{d}r^2}+V(r)+\frac{\left[J(J+1)-\Lambda^2\right]\hbar^2}{2\mu r^2}\right\}\varphi(r)\\&=E_{\upsilon J}\varphi(r),\end{aligned}\tag{2}$$

其中υ和J分别是振动和转动量子数, Λ是电子轨道角动量在在原子核线上的投影 (由电子态决定), μ是两个原子核的约化质量, $\hbar=h/(2\pi)$(h是普朗克常数). 由方程 (2) 可以看出, 振动能级的求解依赖于势能曲线 $V(r)$. 由于 CCSD(T) 方法是所有 CC 理论近似方法中的“黄金标准”[20], 因此本文采用 CCSD(T)/cc-pVXZ 方法获取分子势能. 其中 X 取 T ($T=3$), Q ($Q=4$) 或 5. 在得到势能曲线后, LEVEL 程序[21]被用来实际求解方程 (2), 从而获得分子的振动能级.

1.2 联立机器学习算法和从头算方法

1.2.1 机器学习算法

机器学习算法有两个要素, 一是数据集, 二是学习模型. 机器学习算法通过学习模型, 从数据集提取信息, 以构建输入变量 $X:\{x_1,x_2,\cdots,x_n\}$与输出变量 $Y:\{y_1,y_2,\cdots,y_m\}$之间的映射关系[22]. 为了获得预测的鲁棒性 (即算法在处理存在噪声或离群点的数据时, 仍能保持良好性能的能力[23]), 样本数据集通常会被划分为训练集 (构建高维映射函数) 和测试集 (测试映射函数的预测性能). 在众多的学习算法中, 随机森林回归算法 (random forest regression algorithm, RFRA) 因其适用领域广泛, 超参数数量少, 是常用的监督机器学习算法之一, 已经被成功应用于各个领域[24–31]. 其原理如图 1 所示, 先将训练集再抽样, 分为n个随机树, 即样本子集. 再对随机树j进行分割, 最优分割方案$\theta_j(\theta_1,\theta_2,\cdots,\theta_n$之间彼此独立同分布) 由输出变量$Y$与样本值$\hat{Y}$的均方误差 (或绝对均方误差) 决定. 随机树函数记为$h\left(X;\theta_j\right)$. 根据引入的随机性方式不同, 从极端随机分类 (Breiman[26], Cutler 和 Zhao[27]) 到更复杂的数据相关策略 (Amit 和 Geman[28], Breiman[29], Dietterich[30]), 随机树函数的形式不同[31]. 但是 RFRA 的输出变量Y始终是$h\left(X;\theta_j\right)$的均值, 记为$\bar{h}\left(X\right)$:

$$\bar{h}\left(X\right)=\frac{1}{n}\sum_{j=1}^{n}h\left(X;\theta_j\right).\tag{3}$$

值得一提的是, 当$n\to\infty$时, 大数定理保证了

$$E_X\left(\hat{Y}-\bar{h}\left(X\right)\right)^2\to E_X\left(\hat{Y}-E_\theta\left(h\left(X;\theta_j\right)\right)\right)^2.\tag{4}$$

其中, $E(*)$表示$*$的均值. 这使得 RFRA 不容易出现过度拟合[27], 是其一个突出的优点.

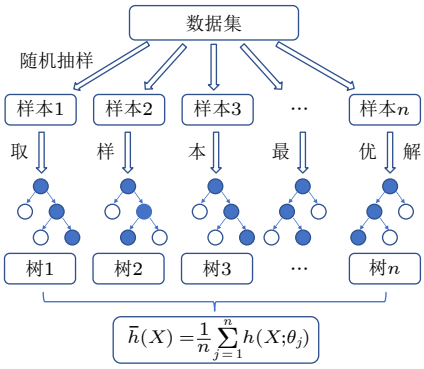

图 1 RFRA 结构图

Fig. 1. Architecture of RFRA.

1.2.2 振动能级预测

由于 CCSD(T) 等从头算方法存在物理简化和数学近似[32], 会导致其与实验值之间存在误差. 不同分子的误差趋势相似 (振动量子数越大, 误差

越大), 但细节不同. 以 CCSD(T)/cc-pVQZ 为例 (见图 2), 可以直观地发现这种趋势. 对于 CCSD(T)/cc-pVTZ 也有类似的结论. 机器学习算法恰好擅长处理这种有大致趋势同时又充满细节的高维度函数映射问题.

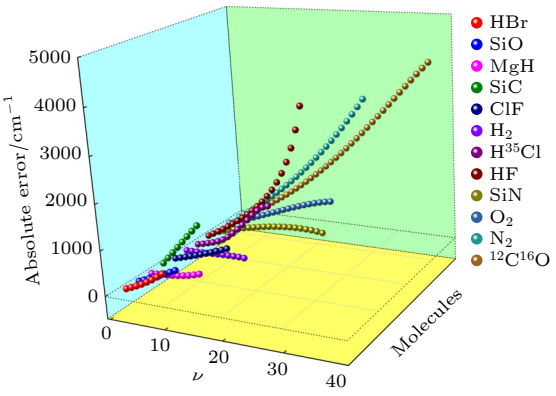

图 2 CCSD(T)/cc-pVQZ 的误差趋势图

Fig. 2. Error trend figure of CCSD(T)/cc-pVQZ.

因此, 通过 RFRA 学习 CCSD(T) 方法的误差分布模式 $E_\nu^{\rm offset}(\alpha)$, 再以此修正从头算方法的理论计算结果 $E_\nu^{\rm ab}$, 从而得到更为精确的预测值 $E_\nu^{\rm P}(\alpha)$:

$$E_\nu^{\rm P}(\alpha) = E_\nu^{\rm offset}(\alpha) + E_\nu^{\rm ab}. \tag{5}$$

其中 α 表示双原子分子种类.

2 结果与讨论

2.1 从头算方法

通过 Gaussian 09 软件 [33] 获取了大量分子体系的 CCSD(T)/cc-pVTZ 和 CCSD(T)/cc-pVQZ 的势能曲线, 并计算了相应的振动能级 $E_\nu^{\rm ab}$. 通过与对应分子的实验值 [34–58] 比较 (部分数据见表 1), 进一步建立了计算能谱的误差数据集 $\mathcal{E}$. 定义理论值与实验值的相对误差为 δ:

$$\delta = \left|\frac{E_\nu^* - E_\nu}{E_\nu}\right| \times 100\%. \tag{6}$$

其中, E_ν^* 代表理论值, E_ν 代表实验值. RFRA 的输入数据分为输入变量 X 和目标变量 $\hat{Y}$ (例见表 2 的 Li_2), 其中 X 包括:

1)CCSD(T)/cc-pVTZ 的振动能级: $E_\nu^{\rm T}$;

2)CCSD(T)/cc-pVQZ 的振动能级: $E_\nu^{\rm Q}$.

RFRA 的任务就是构建这些变量与系统偏差之间的映射关系:

$$(E_\nu^{\rm T}, E_\nu^{\rm Q}) \mapsto E_\nu^{\rm offset}. \tag{7}$$

对比表 2 中 Li_2 的 CCSD(T)/cc - pVQZ 的振动能级和文献 [59] 改进后的振动能级, 可以发现两者在各处的振动能级都非常接近. 并且随着振动量子数增加, 振动能级增大趋势相同. 进一步对比表 2 中的本文预测能级和文献 [59], 可以发现两者各有优劣, 前者在一些能级点的误差更小, 例如 $\nu = 20$ 时, 本文误差为 0.135 cm^{-1}, 而文献 [59] 的误差为 38.701 cm^{-1}. 这说明本文的 CCSD(T) 方法计算结果可信.

2.2 预测结果

将数据集 $\mathcal{E}$ 分为训练集 (38 个分子) 和测试集 (4 个分子) 两部分. 每次预测需要反复训练算法 20 次, 以 RFRA 得到一次测试集结果为一次训练, 最终测试集预测值为每次训练值的均值.

RFRA 的预测性能由相对误差 (参见 (6) 式) 评定. 以 HF, HBr, $H^{35}Cl$ 和 $Na^{35}Cl$ 为例, 预测结果的对比如下 (详见表 3). 训练集的平均相对误差为 0.785%, 最大相对误差为 12.471%, 最小相对误差为 0.000548%; 测试集的平均相对误差为 1.136%, 最大相对误差为 6.601%, 最小相对误差为 0.00964%.

2.3 比较与分析

新方法预测结果误差与更复杂的纯从头算方法 (CCSD(T)/cc-pV5 Z) 的误差对比情况见图 3. 从图 3 可以看出, 结合了机器学习算法 RFRA 与从头算 CCSD(T) 方法的预测能级精度要显著优于 CCSD(T)/cc-pV5 Z, 并且新方法在高能级区仍能保持较好的精度. 新方法的预测能级误差在绝大部分情况下都比 CCSD(T) 的误差小, 只是在振动量子数较低时, 由于算法可能学习到了误差中的随机信息, 所以存在个别点的预测性能下降. 值得注意的是, 当振动量子数较大时, 新方法的优势大幅提升. 例如, $H^{35}Cl$, $\nu = 12$ 时, CCSD(T) 的误差超过了 1000 cm^{-1}, 而新方法的误差仅在 0 附近. 当然, 形式上, 也存在其他利用 CCSD(T)/cc-pVTZ 和 CCSD(T)/cc-pVTZ 的结果进行外推预测的算法, 例如 CCSD(T)/CBS. 本文利用 CCSD(T)/CBS 方法计算了 HF, HBr, $H^{35}Cl$ 和 $Na^{35}Cl$ 的能谱, 其平均相对误差依次为 1.830%, 4.090%, 7.489% 和

表 1 基态分子的部分实验振动能级 (cm^{-1})

Table 1. Partial experimental vibrational energy levels of molecules in the ground state (cm^{-1}).

ν	H_2	HF	DF	$H^{35}Cl$	LiH	Li_2
0	2179.69	2050.77	1490.30	1483.88	705.08	175.032
1	6341.75	6012.19	4396.97	4369.86	2078.45	521.26
2	10268.40	9801.57	7212.12	7151.86	3406.08	862.26
3	13694.54	13423.60	9937.66	9830.66	4688.81	1198.00
4	17433.22	16882.45	12575.33	12406.71	5927.55	1528.41
	$H^{79}Br$	N_2	BF	ClF	NO	CN
0	1314.65	1175.77	742.00	191.77	948.50	1031.20
1	3873.57	3505.69	2208.00	573.06	2824.50	3073.60
2	6341.99	5806.93	3651.00	951.35	4627.30	5089.70
3	8719.91	8079.47	5072.00	1326.60	6491.90	7079.50
4	11007.01	10323.28	6470.00	1698.90	8283.50	9042.80
	Br_2	BCl	CP	CS	SiCl	O_2
0	162.38	416.83	618.19	709.00	267.25	664.69
1	485.53	1242.63	1844.32	2117.00	798.54	2326.53
2	806.51	2058.80	3056.76	3515.00	1325.50	3968.09
3	1125.29	2865.49	4255.53	4902.00	1847.50	5589.60
4	1441.88	3662.85	5440.62	6278.00	2365.50	7191.32
	$^{12}C^{16}O$	$^{13}C^{16}O$	$^{14}C^{16}O$	$^{12}C^{17}O$	$^{13}C^{17}O$	$^{14}C^{17}O$
0	1081.77	1057.72	1036.74	1068.03	1043.66	1022.39
1	3225.04	3153.79	3091.61	3184.32	3112.11	3049.05
2	6341.83	5224.54	5122.15	5274.81	5155.92	5052.07
3	7432.21	7270.04	7128.44	7339.54	7175.14	7031.50
4	9496.24	9290.35	9110.52	9378.60	9169.83	8987.38
	SO	SiC	SiN	$^{24}Mg^{16}O$	$Na^{35}Cl$	NaLi
0	576.94	475.47	574.06	391.14	181.90	127.83
1	1740.42	1416.67	1712.46	1165.88	543.05	381.22
2	2916.75	2344.87	2837.85	1930.32	900.70	631.10
3	4105.98	3260.07	3950.20	2684.16	1254.89	877.79
4	5308.16	4162.27	5049.47	3427.11	1605.65	1121.12
	MgH	AlO	AlF	Al Cl	SiO	BH
0	739.11	488.00	394.60	246.60	619.20	1171.06
1	2171.09	1453.40	1180.00	734.10	1848.90	3440.30
2	3539.79	2404.76	1956.10	1217.30	3066.50	5614.11
3	4841.14	3342.19	2722.30	1698.00	4272.30	7694.16
4	6070.50	4265.42	3480.20	2175.20	5466.10	9684.16

4.004%, 而本文方法的平均相对误差依次为 1.468%, 1.955%, 0.899% 和 0.761%. 可见, 新方法全面优于 CCSD(T)/CBS 方法.

另一方面, 新方法大幅度地降低了计算资源消耗, 以 $Na^{35}Cl$ 为例, 新方法时间不到 5 h(包括计算 CCSD(T)/cc-pVTZ 和 CCSD(T)/cc-pVQZ 方法, 单算法预测时间不到 5 min), 而 CCSD(T)/cc-pV5 Z 方法在相同的设备上计算却需 45 h.

为了进一步验证新方法的可靠性, 本文对更为广泛的分子体系进行了预测, 并将其中 11 种不同类型分子的平均相对误差汇总于图 4. 其中蓝色柱子高度是新方法的平均相对误差, 红色柱子高度是 CCSD(T)/cc-pV5 Z 的平均相对误差. 可见, 新方法有很好的分子体系适应性.

表 2 Li_2 的输入数据与预测数据 (cm^{-1})

Table 2. Input data and predicted data of Li_2 (cm^{-1}).

输入变量 X		目标变量 $\hat{Y}$: E_ν	文献[59]	$E_\nu^{\text{this work}}$
E_ν^{T}	E_ν^{Q}			
171.702	172.1623	175.03	174.059	178.903
511.410	513.631	521.26	518.255	531.853
846.059	849.982	862.26	857.832	878.969
1175.599	1181.177	1198.00	1191.397	1173.096
1499.987	1507.169	1528.41	1519.330	1496.133
1819.161	1827.906	1853.46	1841.991	1906.993
2133.063	2143.328	2173.07	2159.852	2106.362
2441.620	2453.368	2487.19	2471.991	2438.853
2744.753	2757.954	2795.74	2778.334	2693.961
3042.368	3057.002	3090.64	3072.703	3102.89
3334.377	3350.421	3395.80	3373.920	3369.871
3620.669	3638.108	3687.11	3663.159	3630.759
3901.126	3919.951	3972.43	3946.618	4009.308
4175.619	4195.927	4251.73	4223.997	4297.662
4444.003	4465.597	4524.78	4495.270	4487.998
4706.126	4729.113	4791.43	4760.610	4530.406
4961.805	4986.210	5051.53	5018.559	4962.152
5210.858	5236.706	5304.93	5270.268	5225.061
5453.074	5480.402	5551.4	5514.831	5506.045
5688.224	5717.081	6790.71	5753.539	5701.189
5916.059	5946.503	6022.66	5983.959	6022.425

注: 文献[59]的数据和 $E_\nu^{\text{this work}}$ 不作为输入数据.

表 3 HF/HBr/$H^{35}Cl$/$Na^{35}Cl$ 的预测振动能级 (cm^{-1})

Table 3. Predicted vibrational energy levels of HF/HBr/$H^{35}Cl$/$Na^{35}Cl$ (cm^{-1}).

HF ν	E_ν	$\delta_{\text{this work}}$	$E_\nu^{\text{this work}}$	δ_ν^{Q}	E_ν^{Q}
0	2050.771	0.04093	2134.714	0.01200	2075.378
1	6012.194	0.05485	6341.983	0.01562	6106.099
2	9801.566	0.03044	10099.884	0.01694	9967.644
3	13423.603	0.001166	13407.950	0.01797	13664.780
4	16882.448	0.004083	16951.383	0.01893	17202.024
5	20181.824	0.0005710	20193.347	0.01990	20583.515
6	23324.620	0.008118	23135.267	0.02093	23812.920
7	26313.146	0.01637	25882.423	0.02205	26893.357
8	29148.927	0.01435	28730.690	0.02327	29827.322
9	31832.367	0.01963	31207.435	0.02464	32616.648
10	34362.909	0.004397	34211.806	0.02618	35262.441
11	36738.405	0.01714	36108.527	0.02794	37764.995
12	38954.943	0.01769	38265.658	0.03000	40123.716
13	41006.593	0.01487	40396.799	0.03244	42337.026
14	42884.443	0.01483	42248.544	0.03539	44402.227
15	44576.005	0.01264	44012.508	0.03902	46315.354

表 3（续） HF/HBr/$H^{35}Cl$/$Na^{35}Cl$ 的预测振动能级（cm^{-1}）

Table 3 (continued). Predicted vibrational energy levels of HF/HBr/$H^{35}Cl$/$Na^{35}Cl$ (cm^{-1}).

HF ν	E_ν	$\delta_{\text{this work}}$	$E_\nu^{\text{this work}}$	δ_ν^{Q}	E_ν^{Q}
16	46064.207	0.009428	45629.921	0.04357	48071.083
17	47325.663	0.006234	47030.644	0.04938	49662.648
18	48328.541	0.0003463	48311.807	0.05697	51081.932
19	49026.508	0.005459	49294.146	0.06717	52319.807
HBr ν	E_ν	$\delta_{\text{this work}}$	$E_\nu^{\text{this work}}$	δ_ν^{Q}	E_ν^{Q}
0	1314.653	0.02080	1341.993	0.01314	1331.929
1	3873.566	0.06601	4129.243	0.01762	3941.805
2	6341.990	0.01130	6413.868	0.01935	6464.682
3	8719.913	0.01104	8816.207	0.02093	8902.436
4	11007.012	0.01316	11151.888	0.02259	11255.684
5	13202.585	0.01387	13385.700	0.02437	13524.341
6	15305.471	0.01795	15580.146	0.02630	15707.958
7	17313.970	0.002268	17274.704	0.02841	17805.806
$H^{35}Cl$ ν	E_ν	$\delta_{\text{this work}}$	$E_\nu^{\text{this work}}$	δ_ν^{Q}	E_ν^{Q}
0	1483.881	0.01052	1468.276	0.01050	1468.294
1	4369.857	0.01009	4413.952	0.003299	4384.273
2	7151.864	0.009209	7217.724	0.006637	7199.330
3	9830.658	0.01910	10018.462	0.008566	9914.869
4	12406.710	0.004748	12465.616	0.01028	12534.187
5	14880.156	0.002706	14839.892	0.01250	15066.200
6	17250.746	0.002461	17208.284	0.01563	17520.446
7	19517.778	0.003098	19457.320	0.01906	19889.759
8	21680.003	0.01529	21348.516	0.02231	22163.716
9	23735.517	0.01306	23425.579	0.02561	24343.381
10	25681.608	0.005119	25813.072	0.02871	26418.895
11	27514.609	0.009392	27256.180	0.03154	28382.336
12	29229.647	9.641 E-05	29226.829	0.03378	30217.072
13	30820.291	0.008591	30555.509	0.03507	31901.224
14	32278.144	0.004052	32408.930	0.03476	33400.117
$Na^{35}Cl$ ν	E_ν	$\delta_{\text{this work}}$	$E_\nu^{\text{this work}}$	δ_ν^{Q}	E_ν^{Q}
0	181.899	0.008954	180.270	0.04204	174.252
1	543.050	0.01550	534.633	0.03299	525.138
2	900.702	0.007215	894.204	0.03107	872.718
3	1254.890	0.01144	1240.538	0.03014	1217.063
4	1605.649	0.005573	1596.700	0.02955	1558.204
5	1953.013	0.008120	1937.155	0.02912	1896.150
6	2297.016	0.005233	2284.995	0.02877	2230.937
7	2637.692	0.001172	2640.784	0.02847	2562.592
8	2975.075	0.01092	2942.574	0.02821	2891.142
9	3309.198	0.001195	3305.243	0.02798	3216.614
10	3640.093	0.01084	3600.653	0.02776	3539.057
11	3967.795	0.001966	3959.992	0.02755	3858.485
12	4292.334	0.002171	4283.017	0.02735	4174.918

表 3（续） HF/HBr/H^{35}Cl/Na^{35}Cl 的预测振动能级 (cm^{-1})

Table 3 (续). Predicted vibrational energy levels of HF/HBr/H^{35}Cl/Na^{35}Cl (cm^{-1}).

Na^{35}Cl ν	E_ν	$\delta_{\text{this work}}$	$E_\nu^{\text{this work}}$	δ_ν^{Q}	E_ν^{Q}
13	4613.743	0.01077	4564.070	0.02717	4488.398
14	4932.054	0.002617	4919.147	0.02699	4798.950
15	5247.298	0.01274	5180.428	0.02681	5106.596
16	5559.506	0.009977	5504.042	0.02665	5411.364
17	5868.710	0.02119	5744.339	0.02648	5713.289
18	6174.940	0.003747	6198.078	0.02632	6012.398
19	6478.226	0.007659	6428.612	0.02617	6308.719
20	6778.599	0.0008396	6772.914	0.02601	6602.272

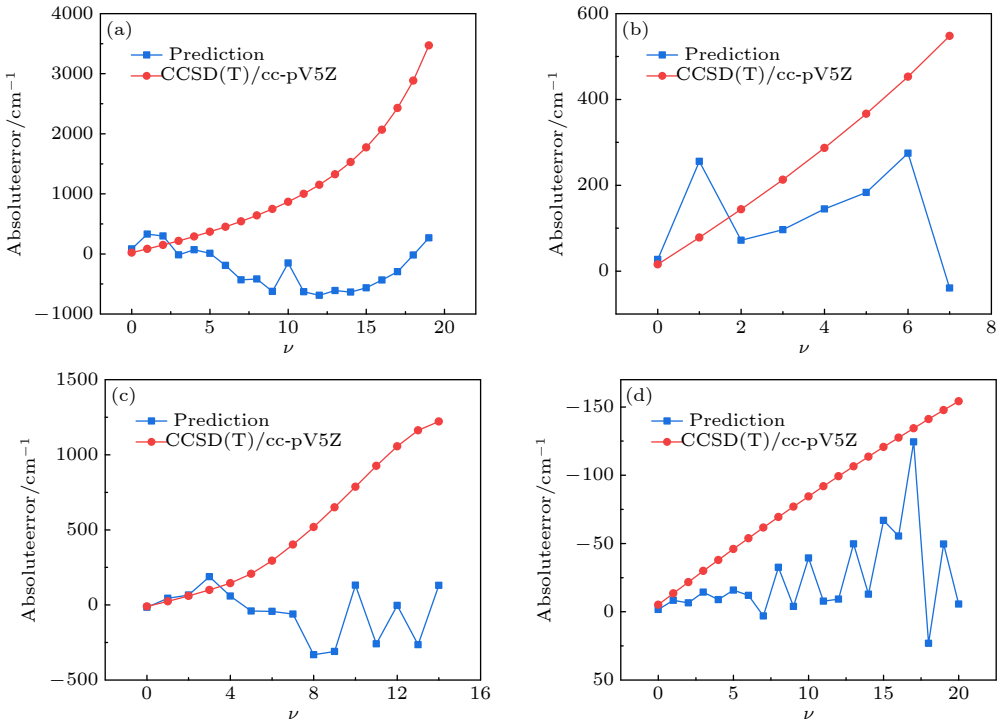

图 3 新方法的预测能级误差与 CCSD(T)/cc-pV5 Z 的误差对比 (a) HF; (b) HBr; (c) H^{35}Cl; (d) Na^{35}Cl

Fig. 3. Comparison of the errors of the new method and CCSD(T)/cc-pV5 Z: (a) HF; (b) HBr; (c) H^{35}Cl; (d) Na^{35}Cl.

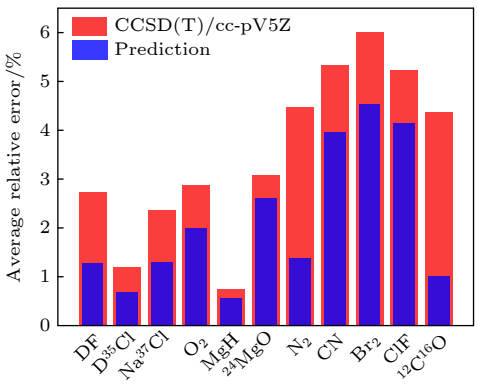

图 4 新方法与 CCSD(T)/cc-pV5 Z 的平均相对误差对比

Fig. 4. Comparison of the average relative errors of the new method and CCSD(T)/cc-pV5 Z.

3 结 论

CC 等从头算方法是获得双原子分子振动能级的常用标准方法. 然而, 由于物理近似和计算量的原因, 从头算方法的预测性能还存在着提升空间. 本文借助机器学习算法 (RFRA), 构建出了从头算方法的误差分布模型, 用于修正能级预测结果. 针对 HF, HBr, H^{35}Cl 和 Na^{35}Cl 等卤素分子的研究表明, 新方法的振动能级预测效果系统地优于纯 CCSD(T)/cc-pV5Z 方法. 新方法将平均相对误差从 2.473% 降到了 1.136%. 此外, 新方法还大幅节约了计算资源, 将典型体系的计算量降低到了原来

的九分之一.

研究结果表明, 机器学习算法可以利用典型从头算方法 (例如 CCSD(T)) 的计算过程信息以及其他类似体系计算的结果, 从大量数据中, 学习到系统性误差的特征, 并构建相应的高维分布函数进行修正, 从而将原始从头算方法的预测误差减小一半, 并且将同一量级计算精度需求下的计算时间缩短一个量级. 值得注意的是, 本文提出的误差分布建模方法, 不仅仅局限于双核体系的能级预测, 在其他从头算方法可以获得数据的领域 (比如大分子体系的能谱性质), 只要有足够的实验数据, 都可以按照相同的模式处理.

参考文献

[1] Ye Y W, Jiang Z L, Zou Y J, Chen H, Guo S D, Yang Q M, Chen L Y 2020 *J. Mater. Sci. Technol.* **43** 144
[2] Wick C D 2017 *J. Chem. Phys.* **147** 161703
[3] Devlin J P, Farnik M, Suhm M A, Buch V 2005 *J. Phys. Chem. A.* **109** 955
[4] Delval C, Fluckiger B, Rossi M J 2003 *Atmos. Chem. Phys.* **3** 1131
[5] Barone S B, Zondlo M A, Tolbert M A 1999 *J. Phys. Chem. A.* **103** 9717
[6] Smart R S C, Sheppard N 1971 *Proc. R. Soc. Lond. A.* **320** 417
[7] Blass P M, Jackson R C, Polanyi J C, Weiss H 1991 *J. Chem. Phys.* **94** 7003
[8] Giorgi J B, Kuhnemuth R, Polanyi J C, Wang J X 1997 *J. Chem. Phys.* **106** 3129
[9] Carvalho A, Hancock G, Saunders M, 2006 *Phys. Chem. Chem. Phys.* **8** 4337
[10] Sun Q 2012 *Vib. Spectros.* **62** 110
[11] Weiss P S, Mestdagh J M, Covinsky M H, Balko B A, Lee Y T 1988 *Chem. Phys.* **126** 93
[12] Rubio L, Samoudi B, Santos M, Diaz L 2012 *J. Photoch. Photobio. A.* **237** 1
[13] Reiser C, Lussier F M, Jensen C C, Steinfeld J I 1979 *J. Am. Chem. Soc.* **101** 350
[14] Rauhut G, Knizia G, Werner H J 2009 *J. Chem. Phys.* **130** 054105
[15] Neff M, Hrenar T, Oschetzki D, Rauhut G 2011 *J. Chem. Phys.* **134** 064105
[16] Kowalski K, Piecuch P 2000 *J. Chem. Phys.* **113** 18
[17] Zhang Y, Zhang Y B, Chen L 2021 *Acta Phys. Sin.* **70** 168702 (in Chinese) [张瑶, 张云波, 陈立 2021 物理学报 **70** 168702]
[18] Nan H, Ma X J, Zhao H B, Tang S J, Liu W H, Wang D W, Jia C L 2021 *Acta Phys. Sin.* **70** 076803 (in Chinese) [南虎, 麻晓晶, 赵海博, 汤少杰, 刘卫华, 王大威, 贾春林 2021 物理学报 **70** 076803]
[19] Li W, Long L C, Liu J Y, Yang Y 2022 *Acta Phys. Sin.* **71** 060202 (in Chinese) [黎威, 龙连春, 刘静毅, 杨洋 2022 物理学报 **71** 060202]
[20] Řezáč J, Šimová L, Hobza P 2013 *J. Chem. Theory Comput.* **9** 364
[21] Le Roy R J. 2017 *J. Quant. Spectrosc. Ra.* **186** 167
[22] Goodfellow I, Bengio Y, Courville A 2016 *Deep Learning* (Cambridge: MIT Press) p351
[23] Daszykowski M, Kaczmarek K, Heyden Y V, Walczak B 2007 *Chemom. Intell. Lab. Syst.* **85** 203
[24] Li Y, Zou C F, Maitane B, et al. 2018 *Appl. Energ.* **232** 197
[25] Abdel-Rahman E M, Ahmed F B, Ismail R 2013 *Int. J. Remote Sens.* **34** 712
[26] Breiman L 2000 *Some Infinity Theory for Predictor Ensembles* Technical Report 579 Statistics Dept. UCB
[27] Cutler A, Zhao G 2001 *Comput. Sci. Stat.* **33** 90
[28] Yali A, Donald G 1997 *Neural Comput.* **9** 1545
[29] Leo B 2001 *Mach. Learn.* **45** 5
[30] Dietterich T G 2000 *Mach. Learn.* **40** 139
[31] Biau G, Devroye L 2010 *J. Multivariate Anal.* **101** 2499
[32] Pokluda J, Cerny M, Sob M, Umeno Y 2015 *Prog. Mater.* **73** 127
[33] Frisch M J, Trucks G W, Schlegel H B, et al. J 2016 *Gaussian 09 (Revision A. 02)* Gaussian Inc
[34] Ram R S, Dulick M, Guo B, Zhang K Q, Bernath P F 1997 *J. Mol. Spectrosc.* **183** 360
[35] Saksena M D, Deo M N, Sunanda K, Behere S H, Londhe C T 2008 *J. Mol. Spectrosc.* **247** 47
[36] Shayesteh A, Henderson R D E, Le Roy R J 2007 *J. Phys. Chem. A* **111** 1249
[37] Coxon J A, Hajigeorgiou P G 1990 *J. Mol. Spectrosc.* **142** 254
[38] Coxon J A, Hajigeorgiou P G 2015 *J. Quant. Spectrosc. Radiat. Transf.* **151** 133
[39] Blom C E, Hedderich H G, Lovas F J, Suenram R D, Maki A G 1992 *J. Mol. Spectrosc.* **152** 109
[40] Edwards S, Roncin J Y, Launay F, Rostas F 1993 *J. Mol. Spectrosc.* **162** 257
[41] Slanger T G, Cosby P C. 1988 *J. Phys. Chem.* **92** 267
[42] Le Roy R J, Appadoo D R T, Colin R, Bernath P F 2006 *J. Mol. Spectrosc.* **236** 178
[43] Reddy R R, Rao T V R, Viswanath R, Viswanath R 1992 *Astrophys. Space Sci.* **189** 29
[44] Speth R S, Braatz C, Tiemann E 1998 *J. Mol. Spectrosc.* **192** 69
[45] Reddy R R, Nazeer A Y, Rama G K, Azeem P A, Anjaneyulu S 1998 *Astrophys. Space Sci.* **262** 223
[46] Venkataramanaiah M, Lakshman S V J 1981 *J. Quant. Spectrosc. Radiat. Transfer* **26** 11
[47] Cai Z L, Martin J M L, François J P, Gijbels R 1996 *Chem. Phys. Lett.* **252** 398
[48] Botschwina P 1986 *J. Mole. Spectrosc.* **118** 76
[49] Dabrowski I 1984 *Can. J. Phys.* **62** 1639
[50] Fellows C E 1991 *J. Chem. Phys.* **94** 5855
[51] Coxon J A, Hajigeorgiou P G 2000 *J. Mol. Spectrosco.* **203** 49
[52] Focsa C, Li H, Bernath P H 2000 *J. Mol. Spectrosco.* **200** 104
[53] Fallon R J, Vanderslice J T, Cloney R D 1962 *J. Chem. Phys.* **37** 1097
[54] Reddy R R, Ahammed Y N, Basha D B, Narasimhulu K, Reddy S S, Gopa K R 2006 *J. Quant. Spectrosco Radiat. Transf.* **97** 344
[55] Barakat B, Bacis R, Carrot F, Churassy S, Crozet P, Martin F, Verges J 1986 *Chem. Phys.* **102** 215
[56] Peterson K A, Woods R C 1987 *J. Chem. Phys.* **87** 4409
[57] Clyne M A A, McDermid I S 1976 *J. Chem. Soc.* **72** 2242
[58] Coxon J A, Hajigeorgiou P G 1992 *Can. J. Phys.* **70** 40
[59] Shi D H, Sun J F, Zhu Z L, Ma H, Yang X D 2008 *Acta Phys. Sin.* **57** 165 (in Chinese) [施德恒, 孙金锋, 朱遵略, 马恒, 杨向东 2008 物理学报 **57** 165]

Combining machine learning algorithm to improve prediction performance of ab initio method for vibrational energy spectra of HF/HBr/$H^{35}Cl$/$Na^{35}Cl$*

Yang Zhang-Zhang [1)] Liu Li [1)] Wan Zhi-Tao [1)] Fu Jia [1)†] Fan Qun-Chao [1)‡]
Xie Feng [2)] Zhang Yi [3)] Ma Jie [4)]

1) (*School of Science, Key Laboratory of High Performance Scientific Computation, Xihua University, Chengdu*, 610039, *China*)

2) (*Institute of Nuclear and New Energy Technology, Collaborative Innovation Center of Advanced Nuclear Energy Technology, Key Laboratory of Advanced Reactor Engineering and Safety of Ministry of Education, Tsinghua University, Beijing*, 100084, *China*)

3) (*College of Advanced Interdisciplinary Studies, National University of Defense Technology, Changsha* 410073, *China*)

4) (*State Key Laboratory of Quantum Optics and Quantum Optics Devices, Laser Spectroscopy Laboratory, College of Physics and Electronics Engineering, Shanxi University, Taiyuan* 030006, *China*)

(Received 13 October 2022; revised manuscript received 29 December 2022)

Abstract

Halides play an important role in atmospheric chemistry, corrosion of steel, and also in controlling the abundance of O_3. Moreover high-precision vibrational energy spectra contain a large amount of quantum information of molecular system and are basic data for people to understand and manipulate molecules. At present, ab-initio methods have achieved many calculation results of the potential energy surfaces and corresponding vibrational energy of molecules, but they still face challenges in terms of accuracy and computational cost. Recently, data-driven machine learning methods have demonstrated very strong capability of extracting high-dimensional functional relationships from massive data and have been widely used in spectrum studies. Therefore, a theoretical approach to combining ab-initio method and machine learning algorithm is presented here to predict the vibrational energy of diatomic systems, which improves the accuracy and simultaneously reduces the computational cost. Firstly, the vibrational energy levels of 42 diatomic molecules are obtained by using different CCSD(T) methods to calculate the configurations from simple to complex and the corresponding experimental results are also collected. A machine learning algorithm is then used to learn the difference between the CCSD(T) method calculated vibrational results and the experimental vibrational results, and a high-dimensional error function is finally constructed to improve the original CCSD(T) computational accuracy. The results for HF, HBr, $H^{35}Cl$ and $Na^{35}Cl$ (they did not appear in the training set) and other halogen molecules show that compared with the CCSD(T)/cc-pV5Z calculation method alone, the present method reduces the prediction error by more than 50% and the computational cost by nearly one order of magnitude. It is worth noting that the method proposed in this paper is not only limited to the energy level prediction of diatomic systems, but also applicable in other fields where data can be obtained by ab initio methods and experimental methods simultaneously, such as the energy spectrum properties of macromolecular systems.

Keywords: vibrational spectrum, ab initio, machine learning, halogen molecules

PACS: 31.15.A–, 33.15.Mt, 33.20.–t, 33.20.Vq **DOI:** 10.7498/aps.72.20221953

* Project supported by the National Natural Science Foundation of China (Grant No. 11904295), the Program of Science and Technology of Sichuan Province of China (Grant No. 2021ZYD0050), and the Open Research Found Program of Collaborative Innovation Center of Extreme Optics (Grant No. KF2020003).

† Corresponding author. E-mail: fujiayouxiang@126.com

‡ Corresponding author. E-mail: fanqunchao@sina.com

物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

结合机器学习算法提高从头算方法对HF/HBr/$H^{35}Cl$/$Na^{35}Cl$振动能谱的预测性能

杨章章 刘丽 万致涛 付佳 樊群超 谢锋 张燚 马杰

Combining machine learning algorithm to improve prediction performance of ab initio method for vibrational energy spectra of HF/HBr/$H^{35}Cl$/$Na^{35}Cl$

Yang Zhang-Zhang Liu Li Wan Zhi-Tao Fu Jia Fan Qun-Chao Xie Feng Zhang Yi Ma Jie

引用信息 Citation: Acta Physica Sinica, 72, 073101 (2023) DOI: 10.7498/aps.72.20221953
在线阅读 View online: https://doi.org/10.7498/aps.72.20221953
当期内容 View table of contents: http://wulixb.iphy.ac.cn
