

Coulombic atomic descriptor for machine learning applications in condensed matter physics

Akram Zarandi, Ali Sadeghi*

Department of Physics, Shahid Beheshti University, Tehran, Iran

Received: 14.04.2019 Final revised: 31.07.2019 Accepted: 28.10.2019

Doi: [10.22055/JRMBS.2019.14920](https://doi.org/10.22055/JRMBS.2019.14920)

Abstract

A main class of machine learning approaches aims at predicting a label or value of some quantity from a set of input data (e.g., recognizing a face from the pixels of a digital image). As an example of the application of such techniques in computational condensed matter physics, we demonstrate, in this study, an accurate prediction of the atomic contributions into a given physical quantity from the arrangement of neighboring atoms. We introduce a descriptor that quantifies the environment of each atom and is filled by the eigenvalues of an approximate Coulomb matrix. The descriptor is invariant under rotation or translation of the molecule and the permutation of the atomic indices. It captures fine structural deformations including the change of the four-body, dihedral angles. Employing this atomic descriptor, we exemplify a promising case where the charges on different atomic species in the molecule are predicted by machine learning to within one tenth of the elementary charge.

Keywords: environment descriptor, machine learning, atomic charge, Coulomb matrix

*Corresponding Author: ali_sadeghi@sbu.ac.ir



توصیف گرمی کولنی برای کاربست در یادگیری ماشین در ماده چگال

اکرم زرنندی، علی صادقی*

گروه فیزیک سامانه‌های پیچیده و زیستی، دانشکده فیزیک، دانشگاه شهید بهشتی، تهران، ایران

دریافت: 1398/01/25 ویرایش نهایی: 1398/05/09 پذیرش: 1398/08/06

Doi: [10.22055/JRMBS.2019.14920](https://doi.org/10.22055/JRMBS.2019.14920)

چکیده

هدف دسته‌ای مهم از رهیافت‌های یادگیری ماشین، پیش‌بینی یک برچسب یا مقدار یک کمیت بر اساس مجموعه‌ای از داده‌های ورودی است (مثل تشخیص دادن یک چهره در پیکسل‌های یک تصویر). به‌عنوان نمونه‌ای از کاربرد چنین روش‌هایی در فیزیک ماده چگال محاسباتی، نشان می‌دهیم که چگونه می‌توان سهم‌های اتمی از یک کمیت فیزیکی را بر مبنای آرایش همسایه‌های آن اتم پیش‌بینی کرد. برای کمی کردن محیط پیرامون یک اتم، توصیف‌گری معرفی می‌کنیم که از طیف ویژه‌مقادیر ماتریس تقریبی کولن ساخته می‌شود. این توصیف‌گر نسبت به چرخش یا انتقال صلب مولکول و نیز جایگشت شماره ترتیب اتم‌های آن ناورداست و تغییرات ظریف ساختاری، از جمله تغییر زاویه دوسطحی که یک کمیت چهارجسمی است، را تشخیص می‌دهد. در قالب یک مثال کاربردی نشان می‌دهیم که با بهره‌گیری از این توصیف‌گر در فرآیند یادگیری، بار الکتریکی روی انواع مختلف اتم‌ها در یک مولکول با خطایی کمتر از یک دهم بار الکترون قابل پیش‌بینی است.

کلیدواژگان: توصیف پیرامون اتم، یادگیری ماشین، بار اتمی، ماتریس کولن

مقدمه

ورودی‌ها و یک ویژگی ماده پیدا می‌شود. تذکر این نکته مفید است که در همه این مثال‌ها بر خلاف شیوه معمول در فیزیک، یعنی یافتن یک مدل تا حد امکان ساده اما قابل توجیه بر مبنای شهود، روابط ریاضی بزرگ و با درجات آزادی بالا برای یافتن الگو یا توصیف ارتباط بین داده‌ها استفاده می‌شود. این امر منجر به چالشی بین منتقدان و طرفداران رهیافت یادگیری ماشین در فیزیک ماده چگال شده است: از یک سو انتقاد برخی فیزیک‌دانان که معتقدند حل مسئله باید اساساً بر مبنای شهود فیزیکی باشد برانگیخته شده است و از سوی دیگر برخی آن را نقطه قوت این رهیافت به‌شمار می‌آورند و معتقدند که ممکن است با کنار گذاشتن قیود ناشی از دخالت شهود فیزیکی نتایج نوینی حاصل شود که تا کنون با اعمال قیود از چشم ما پنهان مانده‌اند.

در سال‌های اخیر یادگیری ماشین به‌حل مسأله‌های پیچیده مهندسی کمک شایانی کرده و در کاربردهای متنوعی، از دسته‌بندی ایمیل‌ها گرفته تا ابراز هویت با عنیبه چشم، موفقیت‌های چشم‌گیری حاصل شده است. هم‌زمان فیزیکدانان محاسباتی هم در این زمینه فعال و موفق بوده‌اند و این‌گونه روش‌ها را برای حل مسائل فیزیک ماده چگال از جمله برای شناسایی گذر فاز مواد [1]، حل معادله شرودینگر بس‌ذره‌ای [2]، تخمین چگالی الکترون در چارچوب نظریه تابعی چگالی اما بدون حل معادلات کوهن-شام [3]، ساختن میدان‌های نیروی بین اتمی [4] و ... به‌کار گرفته‌اند. دو مثال نخست نمونه‌هایی از به‌کار بستن تکنیک‌های بدون نظارت در یادگیری ماشین‌اند درحالی‌که در دو مثال بعدی با استفاده از روش‌های نظارت شده رابطه‌ای بین

* نویسنده مسئول: ali_sadeghi@sbu.ac.ir

باز نشر این مقاله با ذکر منبع آزاد است.

این مقاله تحت مجوز کپی‌رایت کامنز تخصصی 4.0 بین‌المللی می‌باشد



به پیرامونشان به شبکه‌های عصبی مصنوعی آموزش داده و انرژی و نیروها را بر اساس آن تخمین زدند [7]. بدیهی است سهم هر اتم از کمیت‌هایی مانند چگالی الکترون، انرژی یا بار الکتریکی یک سامانه بس‌اتمی وابسته به پیرامون آن (تعداد و نوع اتم‌های همسایه و فاصله و جهت‌گیری آنها) است. بنابراین دقت تخمین این سهم‌ها به‌طور حیاتی به کیفیت توصیف کردن اتم در میان همسایگانش وابسته است. هدف این مقاله معرفی یک توصیف‌گر جدید و کارا برای به‌خدمت گرفتن یادگیری ماشین در حل چنین مسائلی است. توصیف‌گر جدید که برای کاربردهای مبتنی بر جمع‌پذیری در سامانه‌های دارای تنوع بالا در اتم‌های سازنده مناسب است، توسعه‌یافته‌ی توصیف‌گری است که پیشتر توسط نویسندگان و همکاران برای مقایسه‌ی پیکربندی‌های مختلف ساختارهای خوشه‌ای و بلورین معرفی شده بود [8-9].

در ادامه، ابتدا این توصیف‌گر معرفی و ویژگی‌هایی که آن را برای کاربردهای یادگیری ماشین مناسب می‌کند تشریح می‌شود. سپس مثالی از کاربرد آن برای پیش‌بینی بارهای اتمی با یادگیری ماشین ارائه می‌شود.

مدل و فرمول‌بندی

منظور ما از توصیف‌گر یا اثر انگشت یک اتم، آرایه‌ای مرتب از اعداد است که اتم مربوطه را در بین همسایگانش توصیف می‌کند. علی‌الاصول با داشتن این آرایه پیرامون اتم شناسایی و بر همین مبنا سهم اتم از بار، انرژی، چگالی الکترون و یا هر کمیت مورد علاقه دیگر پیش‌بینی می‌شود. به عبارتی دیگر، ماشین آموزش دیده آرایه را به‌عنوان ورودی دریافت می‌کند و کمیت فیزیکی معینی را به‌عنوان خروجی پیش‌بینی می‌کند.

خاصیت جمع‌پذیری، یعنی بسط یک ویژگی کمی انبوه‌ای از اتم‌ها برحسب سهم‌های اجزای سازنده آن، یک مسأله جالب در فیزیک ماده چگال است که با رهیافت یادگیری ماشین قابل بررسی است. مثلاً نظریه تابعی چگالی جزء‌جزئی¹ [5] ایده بیان چگالی الکترون $n(\mathbf{r})$ به‌صورت مجموع سهم‌های جایگزیده N قسمت مجزای یک سامانه را پیشنهاد می‌دهد: $\sum_{i=1}^N n_i(\mathbf{r}) = n(\mathbf{r})$. با این شرط، مجموع انرژی مربوط به همه قسمت‌ها نسبت به سهم‌های جزئی کمینه می‌شود: $\min_{n_i(\mathbf{r})} \sum_{i=1}^N E[n_i(\mathbf{r})]$. از حیث عملیاتی فروکاستن یک کمیت به جمع جبری سهم‌های اجزای سازنده سامانه، کاهش چشم‌گیری در هزینه‌های محاسباتی به‌ارمغان می‌آورد. به‌عنوان مثالی ساده و واقعی فرض کنید انرژی یک سیستم N اتمی را، هر چند به‌تقریب، برحسب انرژی اتم‌های آن بنویسیم $E = \sum_{i=1}^N E_i$. اگر بتوان سهم اتم i از انرژی، E_i را بسته به نوع آن اتم و چیدمان اتم‌های پیرامونش تا شعاعی محدود تعیین کرد هزینه محاسبه انرژی یک سامانه بزرگ فقط به‌نسبت تعداد اتم‌های سازنده آن N زیاد می‌شود. چنین مقیاس‌پذیری خطی نسبت به اندازه سامانه یک حد ایده‌آل در محاسبات مبتنی بر ساختار الکترونی است [6]. در دهه اخیر تلاش‌هایی برای یافتن سهم‌های اتمی در انرژی با استفاده از رهیافت‌های یادگیری ماشین صورت گرفته است. مثلاً بلر و پاریلنو یک مجموعه شبکه عصبی مصنوعی چند لایه را با داده‌های محاسبه شده از نظریه تابعی چگالی آموزش داده و موفق شدند با دقت بالایی انرژی کل و نیروهای وارد بر اتم‌ها را برای طیف وسیعی از سامانه‌های اتمی تخمین بزنند [4]. در روشی متفاوت، قاسمی و همکاران الکترون‌خواهی و بارهای اتم‌ها را با توجه

¹ Partition density functional theory

درایه در یک ماتریس چیده و پس از قطری کردن، طیف ویژه‌مقادیر را به ترتیب نزولی در آرایه توصیف‌گر لیست می‌کنیم. [۸،۹] حال مقایسه محیط دو اتم با مقایسه طیف‌ها امکان‌پذیر می‌شود: هر چه وضع پیرامونی دو اتم تفاوت کمتری داشته باشد (یعنی تعداد، فاصله، جهت‌گیری و نوع همسایه‌های آنها شبیه‌تر باشد) طیف آنها تشابه بیشتری به هم خواهد داشت. کارایی و قدرت تمیز این روش در مقالات قبلی نشان داده شده است. [۸،۹] مانسته فیزیکی و آزمایشگاهی این مقایسه، مطالعه طیف اتمی در آزمایش‌هایی مثل XPS است که نوع و طول پیوندهای یک گونه اتمی در ترکیب از طیف آن اتم شناسایی می‌شود.

در این مقاله نمونه جدیدی از این توصیف‌گر معرفی می‌کنیم که با استفاده از شهود فیزیکی اطلاعات بیشتری از اتم‌ها را در ساخت ماتریس به‌کار می‌گیرد. لذا انتظار می‌رود برای نمونه‌هایی که تنوع گونه‌های اتمی آن بیشتر است کارا تر عمل کند. درایه ij این ماتریس با رابطه

$$C_{ij} = \iint \frac{\rho_i(\mathbf{r})\rho_j(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}d\mathbf{r}' = \frac{q_i q_j}{r_{ij}} \text{erf}(\sqrt{\gamma_{ij}} r_{ij})$$

داده می‌شود که در آن $r_{ij} = \|\mathbf{r}_i - \mathbf{r}_j\|^2$ فاصله بین دو اتم i و j و $\rho_i(\mathbf{r}) = \gamma_{ij} = \alpha_i \alpha_j / (\alpha_i + \alpha_j)$ و

$q_i \left(\frac{\alpha_i}{\pi}\right)^{3/2} e^{-\alpha_i \|\mathbf{r}-\mathbf{r}_i\|^2}$ توزیع گاوسی بار الکتریکی به‌بزرگی q_i و به‌مرکز \mathbf{r}_i یعنی محل اتم i است. نهایتاً پس از ضرب هر درایه ij در عامل وزنی $r_{ik}^{-1} r_{jk}^{-1}$ که اهمیت اتم‌ها را متناسب با فاصله‌شان تا اتم مورد نظر k تضعیف می‌کند، ماتریس را قطری کرده و به‌ترتیبی که قبلاً بیان شد بردار توصیف‌گر اتم k ساخته می‌شود.

از نظر فیزیکی، درایه‌های ماتریس از جنس برهم‌کنش کولنی‌اند. به‌طور دقیق‌تر درایه G_{ij} تقریب مرتبه اول از

برای سادگی توضیحات خود را با یک توصیف‌گر ساده یعنی عدد هم‌آرایی¹ شروع می‌کنیم که پیرامون هر اتم را با یک عدد صحیح (تعداد اولین همسایه‌ها) توصیف می‌کند. دو نقص این توصیف‌گر عبارتند از گسسته بودن (تعلق به مجموعه اعداد صحیح) و لحاظ نکردن همسایه‌های بعدی (از دست رفتن بخش مهمی از اطلاعات مفید). راه حل رفع همزمان این کاستی‌ها بسیار ساده است. کافی است از انتگرال هم‌پوشانی به‌جای عدد هم‌آرایی استفاده کنیم. به‌عبارت واضح‌تر، انتگرال‌های هم‌پوشانی بین یک تابع گاوسی به‌مرکز اتم مورد نظر و توابع گاوسی مشابهی به‌مراکز اتم‌های مجاور (که پهنا و ارتفاع آنها را به‌شکلی که در پایین توضیح داده خواهد شد بسته به‌نوع اتم تنظیم کرده‌ایم) محاسبه و آنها را با نسبت وزنی متناسب با عکس فاصله تا اتم مورد نظر با هم جمع می‌کنیم. این عدد هم‌آرایی تعمیم یافته غیرصحیح، اندکی اثر همسایه‌های دوردست را هم لحاظ کرده است و دارای یک ویژگی مهم و کاربردی است: ناوردایی نسبت به‌انتقال و چرخش فضایی ساختار. ولی این توصیف‌گر صرفاً بر مبنای فواصل دو ذره‌ای ساخته شده و از اطلاعات چند-ذره‌ای و جهت‌گیری اتم‌ها حول اتم مورد نظر بهره‌ای نمی‌برد.

نظریه گراف برای رفع این کاستی راه‌گشاست. به‌عنوان مثال ویژه‌بردار اساسی (یعنی متناظر با بزرگترین ویژه‌مقدار) ماتریس اتصال بین اتمی (با درایه‌های صفر یا یک) که با توجه به‌وجود یا عدم پیوند بین اتم‌ها ساخته می‌شود به‌عنوان اثر انگشت برای تمیز اتم‌ها در ساختار یک مولکول استفاده شده است. [10] ما برای ساختن توصیف‌گر یک اتم در میان اتم‌های همسایه، انتگرال‌های همپوشانی را پس از ضرب در ضریب وزنی متناسب با عکس فاصله تا اتم مورد نظر به‌عنوان

¹ coordination number

مورد آن از حوصله این مقاله خارج است. لذا در اینجا به مشاهده کیفی این تغییرات در مورد یک مولکول نمونه بسنده می‌کنیم و در عوض ارزیابی کمی را بعداً در مورد پیش‌بینی بار الکتریکی انجام خواهیم داد. برای این منظور مولکول آزوبنزن¹ را که از اتصال دو حلقه فنیلی با یک پیوند دوگانه $N=N$ ساخته می‌شود (شکل 1) مورد آزمایش قرار می‌دهیم. چهار نوع تغییر شکل به ترتیب زیر به آن اعمال می‌کنیم:

- 1 180° تغییر زاویه دوسطحی $C-N=N-C$
(پیشش حلقه راست حول محور $N=N$)
- 2 180° تغییر زاویه پیوند $N-N-C$
- 3 180° تغییر زاویه دوسطحی $N=N-C-C$
(پیشش حلقه راست حول محور $C-N$)
- 4 2/5 آنگستروم افزایش فاصله $N=N$

پاسخ چهار مؤلفه توصیف گر کولنی برای دو اتم نیتروژن و یک اتم کربن به این تحولات هندسی نشان می‌دهد که مؤلفه‌های بردار به‌طور مؤثر اتم‌های کربن و نیتروژن را از یکدیگر تمییز می‌دهند. حتی توصیف‌گرهای دو نیتروژن که محیط پیرامونشان به دلیل هم‌صفحه نبودن حلقه‌ها اندکی متفاوت است از یکدیگر تفکیک شده‌اند. توجه کنید که زاویه دوسطحی یک کمیت چهارجسمی است و درک تغییرات آن توسط توصیف گر ما که براساس فواصل بین اتمی (کمیت دوجسمی) ساخته شده است نشان‌دهنده مؤثر بودن همه درایه‌های ماتریس در بردار توصیف گر است. نکته قابل توجه دیگر میزان حساسیت توصیف گر یک اتم به یک تغییر ساختاری است. از دو قسمت آخر نمودار می‌توان دریافت که شدت این حساسیت به نزدیکی یا دوری آن اتم تا محل اعمال تغییر وابسته است.

بسط انرژی برهم‌کنش الکترواستاتیک (هارتری) بین دو اتم i و j است و لذا ماتریس جدید را ماتریس کولن می‌نامیم. استفاده از ماتریس برهم‌کنش الکترواستاتیک بین بارهای نقطه‌ای (یعنی اولین تقریب بسط چندجمله‌ای توزیع بارهای اتمی) که قبلاً پیشنهاد داده شده است [11] با گذاشتن $\gamma_{ij} = 0$ در رابطه فوق که منجر به حذف تابع خطا یعنی $C_{ij} = q_i q_j / r_{ij}$ می‌گردد، حاصل می‌شود. در حالی که در ماتریس کولن معرفی شده در بالا هم‌پوشانی توزیع‌های بار دو اتم i و j از طریق شعاع‌های اتمی آنها یعنی α_i و α_j لحاظ می‌شود. در مقایسه با مرجع 8 که هم‌پوشانی اوربیتال‌ها را مورد استفاده قرار می‌دهد، توصیف گر کولنی بر اساس هم‌پوشانی توزیع‌های گاوسی‌شکل از بارهای اتمی ساخته شده است و لذا بین اتم‌ها با بارهای والانس متفاوت تمییز قائل می‌شود. همچنین توجه کنید که در مرجع 8 یک ساختار به‌طور کلی توصیف می‌شود در حالی که توصیف گر جدید ما هر اتم در ساختار را توصیف می‌کند.

نتایج و بحث

با توجه به کارکردی که از توصیف گر انتظار داریم برای نشان دادن کیفیت و کارایی آن آزمایش زیر را انجام می‌دهیم. ابتدا باید دید که بردار توصیف گر، اتم‌هایی را که وضعیت پیرامونی متفاوتی دارند چگونه از یکدیگر تمییز می‌دهد. سپس با اعمال تغییرات هندسی به ساختار اتمی تحول القا شده در مؤلفه‌های بردار توصیف گر اتم‌هایی که لازم است این تغییر را حس کنند دنبال می‌کنیم تا مطمئن شویم که توصیف گر به اندازه کافی به این تغییرات حساس است. البته می‌توان برای نمایش این تغییرات ملاک‌های کمی هم تعریف کرد که البته تعریف معیار یکتا نیست و بحث مفصل در

¹ Azobenzene

² Dihedral angle

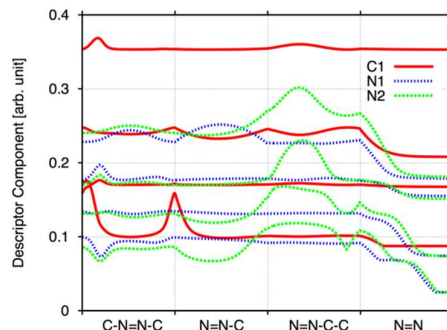
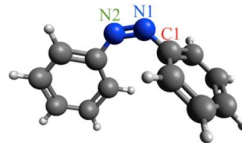
مثال، بردارهای توصیف‌گر به‌عنوان ورودی به‌ماشین داده می‌شود تا بارهای اتمی مولکول را پیش‌بینی کند. ما از روش $1KRR$ که یک تکنیک استاندارد یادگیری ماشین است [12] استفاده می‌کنیم که در آن بر مبنای میزان شباهت به‌داده‌های مرجع، ورودی برداری \mathbf{x} (توصیف‌گر اتمی) به‌خروجی نرده‌ای q (بار اتمی) تصویر می‌شود:

$$q^{KRR}(\mathbf{x}) = \sum_{i=1}^{N_T} \alpha_i e^{-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2\sigma^2}}$$

ضرایب α_i با کمینه کردن تابع هزینه

$$\mathcal{L}(\{\alpha_i\}) = \sum_{i=1}^{N_T} |q^{KRR}(\mathbf{x}_i) - q_i|^2 + \lambda \sum_{i,j} \alpha_i \alpha_j e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}$$

تعیین می‌شوند. جمله اول هزینه در واقع خطای کل پیش‌بینی بار برای N_T داده مرجع موجود در مجموعه آموزشی $\{\mathbf{x}_i, q_i\}$ است. انگیزه افزودن جمله دوم به تابع هزینه جلوگیری از بزرگ شدن ضرایب (که منجر به بیش‌برازش² خواهد شد) است. [12] دو هاپیرپارامتر λ و σ طوری تنظیم می‌شوند که خطای پیش‌بینی برای مجموعه جدیدی از داده‌ها که در مرحله آموزش به‌کار نرفته‌اند کمینه شود (این کار با اعتبارسنجی که در ادامه جزئیات بیشتری در مورد آن بیان خواهد شد انجام می‌شود). پس از این فرآیند که آموزش ماشین نامیده می‌شود، ماشین آماده استفاده است. در آزمایش زیر دقت ماشین در پیش‌بینی بار اتمی را به‌طور کمی می‌سنجیم.



شکل 1. نمای طرح‌واره مولکول آزوبزن و نمودار تحول مؤلفه‌های بردار توصیف‌گر اتمهای C1 و N1 و N2 در پاسخ به چهار مرحله تغییرات هندسی در مولکول: پیش 180 درجه حول محور N=N، تغییر 180 درجه ای زاویه N-N-C، پیش 180 درجه حول محور N-C و افزایش 2/5 آنگسترومی فاصله بین دو نیتروژن (متن را ببینید).

توجه کنید که در اینجا تعداد مولکولی که دارای تعداد کمی درجات آزادی باشد را مثال زده‌ایم تا بتوان اثر هر یک از آنها را جداگانه بررسی کرد. در عین حال دقت شده است که این مولکول دارای تنوع درجات آزادی باشد: طول پیوند، زاویه و زاویه دووجهی (شکل 1). درجات آزادی در هر سیستم دیگری قاعداً یکی از همین انواع است چرا که مبنای کار توصیف‌گر معرفی شده موضعی بودن است و آثار دوربرد ناشی از بزرگ بودن سیستم تأثیری در کار توصیف‌گر نخواهد داشت. هزینه محاسباتی تعیین توصیف‌گر متناسب با تعداد اتم‌ها و شعاع قطع اطراف اتم‌هاست.

از آنجا که هدف نهایی طراحی توصیف‌گر استفاده از آن برای افزایش کارایی فرآیند یادگیری ماشین است، نمونه‌ای از این کاربردها در ادامه ارائه می‌شود. در این

¹ Kernel Ridge Regression

² overfitting

می‌دهد). داشتن طول مناسب بردار توصیف‌گر عامل مهم دیگری در بالابردن کارایی ماشین است. معمولاً با مقایسه خطاها می‌توان به‌سادگی طول مناسبی برای این بردار یافت طوری که پیرامون اتم‌ها را با دقت کافی توصیف کند و در عین حال هزینه محاسباتی را در حد معقولی پایین نگه دارد. این طور پیدا کردن هایپرپارامترها با آزمون و خطا یک ویژگی کلی در همه رهیافت‌های یادگیری ماشین است. ما در این کار بردار توصیف‌گر کولنی 5 بعدی را مناسب یافتیم.

از 500 پیکربندی مولکول 24 اتمی حاضر در مجموع 12000 زوج توصیف‌گر-بار اتمی (نمونه) به‌دست می‌آید. این مجموعه نمونه به‌طور تصادفی به سه دسته و با نسبت 3:1:6 تقسیم می‌شود. دسته اول که دارای $N_T = 3600$ نمونه است برای آموزش و دسته 1200 تایی برای اعتبارسنجی متقابل² به‌خدمت گرفته می‌شود. به‌عبارت دقیقتر، ماشین با نمونه‌های دسته اول آموزش داده می‌شود و سپس خطای پیش‌بینی بار اتمی برای نمونه‌های دسته دیگر محاسبه می‌شود. با پیمایش فضای دوبعدی هایپرپارامتری این کار بارها تکرار می‌شود و بهترین نتایج (یعنی خطای جذر میانگین مربعات³ حدود $0,00015e$ و بزرگترین مقدار خطا کمتر از $0,07e$ که $\lambda = 1.6 \times 10^{-19}C$ بار بنیادی است) به‌ازای 10^{-3} و 10^{-1} حاصل می‌شود. دسته سوم شامل 7200 نمونه برای آزمون نهایی و گزارش نتایج زیر کنار گذاشته شده و در هیچ یک از مراحل آموزش استفاده نمی‌شود. توجه کنید که می‌توان زوج‌های گونه‌های اتمی متفاوت را از هم جدا کرده و برای هر گونه اتمی ماشین جداگانه‌ای آموزش داد. اما تجربه ما در مورد مسأله حاضر آن است که یک ماشین واحد برای همه

برای تولید داده‌ها آزمایشی ترتیب داده‌ایم که در آن مولکول آزوبزن در یک شبیه‌سازی دینامیک مولکولی در دمای ثابت 800 کلون جنب و جوش می‌کند و پیکربندی‌های متفاوتی را تجربه می‌کند. در این حین برای پیکربندی‌های نمونه بارهای اتمی تعیین می‌شود. دینامیک مولکولی و محاسبات اصول اولیه در تقریب چگالی موضعی¹ و با نرم افزار DFTB+ انجام می‌شود. [13] در این نرم‌افزار، تقریبی از نظریه تابعی چگالی پیاده‌سازی شده است که در آن چگالی الکترون برحسب بارهای اتمی بسط داده می‌شود [14]. به‌این ترتیب قسمتی از محاسبات را می‌توان از قبل انجام داد و بعداً به‌صورت پارامتر استفاده کرد. این ویژگی به محاسبات سرعت بسیار بالایی می‌دهد. ما مجموعه پارامترهای مولکول‌های آلی [15] را به‌کار می‌بریم. با حل معادله شروینگر که در پایه اوربیتال‌های گاوسی بیان شده است، بارهای اتم‌ها در یک حلقه خودسازگار تعیین می‌شوند تا کمینه انرژی سیستم حاصل شود. یادآوری می‌کنیم که هدف این قسمت از کار صرفاً نمایش کارایی توصیف‌گر کولنی معرفی شده برای تخمین بارهای اتمی است. اصولاً با هر روش محاسباتی اصول اولیه‌ای (مثلاً با تصویر کردن تابع موج بر اوربیتال‌های اتمی) یافتن بارهای اتمی امکان‌پذیر است. یکی از عوامل تعیین‌کننده در موفقیت یادگیری تنوع کافی در مجموعه داده آموزش است. به‌عبارت دیگر باید نمونه‌برداری از همه فضای پیکربندی انجام شود. برای آنکه مطمئن شویم در 500 پیکربندی نمونه‌برداری شده در این مرحله، نمونه‌های تکراری یا شبیه به هم وجود ندارد از مقایسه بردارهای توصیف‌گر استفاده می‌کنیم. (مرجع 8 به‌طور مفصل شیوه تشخیص یکسان یا متفاوت بودن پیکربندی‌های یک مولکول را شرح

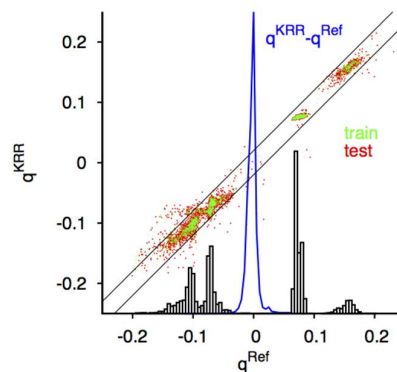
³ root-mean-square error¹ Local Density Approximation (LDA)² cross-validation

چشم‌گیر و نزدیک به ایده‌آل (خط راست نیمساز ربع اول و سوم) است. تعداد نقاط خارج از محدوده‌ای که با دو خط به موازات نیمساز و در فاصله $0,02e$ از آن مشخص شده بسیار اندک است و از بین هزاران نمونه فقط تعداد انگشت‌شماری از خط نیمساز دورند. هیستوگرام بسیار تیز خط $q^{KRR}(x_i) - q_i$ به روشنی دلالت بر آن دارد اغلب نمونه‌ها با خطای ناچیز پیش‌بینی شده‌اند. به علاوه خطای پیش‌بینی برای داده‌های آموزشی و آزمایشی کم و بیش برابر است¹ و برای هر دو مجموعه آموزشی و آزمایش همان مقدار خطای اعتبارسنجی متقابل که در بالا گزارش شد (یعنی خطای جذر میانگین مربعات و بزرگترین خطا به ترتیب $0,00015e$ و $0,07e$) است. این خطا هم مرتبه خطای سیستماتیک روش‌های ابتدا به ساکن در محاسبه بار اتمی است (مثلاً همه هیدروژن‌ها در این مولکول تقریباً مشابه‌اند و لذا باید بار تقریباً یکسانی داشته باشند ولی هیستوگرام بارهای مرجع در شکل 2 برای لکه هیدروژن پهنای حدود $0,02e$ دارد). بر این مبنا می‌توان انتظار داشت که یادگیری ماشین با توصیف‌گر اتمی کولنی می‌تواند جایگزین محاسبات سنگین مبتنی بر اصول اولیه شود تا با دقتی قابل مقایسه با آن، هزینه محاسبات سهم‌های اتمی از بار یا انرژی و ... را به شدت کاهش دهد.

خلاصه و نتیجه‌گیری

در این مقاله نمونه جدیدی از توصیف‌گر اتمی مبتنی بر طیف ویژه‌مقادیر ماتریس کولن معرفی شد. ایده اولیه این توصیف‌گر به نظریه گراف در ریاضیات مربوط است. ماتریس مورد استفاده بر مبنای شهود فیزیکی

گونه‌ها کارایی بهتری دارد و لذا ما به همین روش کار را دنبال می‌کنیم. شکل 2 نمودار پراکندگی بارهای پیش‌بینی شده با ماشین در مقایسه با بارهای اتمی مرجع (محاسبه شده از اصول اولیه) را نشان می‌دهد. قبل از تحلیل آن، به نمودار میله‌ای فراوانی بارهای مرجع توجه کنید که وجود چهار گروه را نشان می‌دهد در حالی که فقط سه گونه اتمی (کربن، نیتروژن و هیدروژن) در مولکول وجود دارد. در واقع اتم‌های کربن دو گروه شده‌اند: برخی بار مثبت و برخی دیگر مانند نیتروژن‌ها بار منفی دارند. لکه کوچک حوالی $0,07e$ متعلق به اتم‌های هیدروژن است. توزیع بارها گستره‌ای به بزرگی حدود $0,5e$ را پوشش می‌دهد که نشانه نمونه‌برداری از بخش‌های مختلف فضای پیکربندی است.



شکل 2. پراکندگی بار الکتریکی پیش‌بینی شده (یادگیری ماشین) نسبت به مقدار مرجع (اصول اولیه). یکای بار، بار بنیادی e است و دو خط موازی حدود $y = x \pm 0,02$ را مشخص کرده‌اند. فراوانی توزیع بارهای مرجع (نمودار میله‌ای) و فراوانی خطاها $q^{KRR} - q^{Ref}$ برای مقایسه بر روی همان محور افقی نشان داده شده‌اند.

حال به خود نمودار پراکندگی توجه می‌کنیم. برای هر چهار دسته، توافق بار پیش‌بینی شده با مقدار مرجع

¹ این برابری نشان دهنده تنظیم مناسب هابیر پارامترها است چرا که اگر بیش-برازش روی دهد خطا برای داده‌های آموزشی کم و برای داده‌های آزمایشی بزرگ خواهد بود.

پژوهشی دانشگاهی از آدرس اینترنتی زیر در دسترس است: <http://comphys.sbu.ac.ir>

مرجع ها

[1] J. Carrasquilla, R.G. Melko, Machine learning phases of matter, *Nature Physics* **13.5** (2017) 431. doi.org/10.1038/nphys4035

[2] G. Carleo, M. Troyer, Solving the quantum many-body problem with artificial neural networks. *Science* **355.6325** (2017) 602-606. doi.org/10.1126/science.aag2302

[3] F. Brockherde, et al., Bypassing the Kohn-Sham equations with machine learning, *Nature communications* **8.1** (2017) 872. doi.org/10.1038/s41467-017-00839-3

[4] J. Behler, M. Parrinello., Generalized neural-network representation of high-dimensional potential-energy surfaces, *Physical review letters* **98.14** (2007) 146401. doi.org/10.1103/PhysRevLett.98.146401

[5] P. Elliott, K. Burke, M.H. Cohen, A. Wasserman, Partition density-functional theory, *Physical Review A* **82.2** (2010) 024501. doi.org/10.1103/PhysRevA.82.024501

[6] S. Goedecker, Linear scaling electronic structure methods, *Reviews of Modern Physics* **71.4** (1999) 1085. doi.org/10.1103/RevModPhys.71.1085

[7] S.A. Ghasemi, A. Hofstetter, S. Saha, S. Goedecker, Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network, *Physical Review B* **92.4** (2015) 045131. doi.org/10.1103/PhysRevB.92.045131

[8] A. Sadeghi, S.A. Ghasemi, B. Schaefer, S. Mohr, M.A. Lill, S. Goedecker, Metrics for measuring distances in configuration spaces, *The Journal of chemical physics* **139**, (2013) 184118. doi.org/10.1063/1.4828704

[9] L. Zhu, M. Amsler, T. Fuhrer, B. Schaefer, S. Faraji, S. Rostami, S.A. Ghasemi, A. Sadeghi, M. Grauzinyte, C. Wolverton, S. Goedecker, A fingerprint based metric for measuring similarities of crystalline structures, *The Journal of chemical physics* **144**, (2016) 034203. doi.org/10.1063/1.4940026

بر اساس برهم کنش کولنی تقریبی بین اتم ها ساخته می شود و گونه اتم های همسایه اتم مورد نظر در مقدار داریه های ماتریس نقش دارد. لذا این توصیف گر برای توصیف پیرامون اتم ها در ترکیبات دارای تنوع گونه اتمی به خوبی قابل استفاده است. چون تعداد دلخواهی از ویژه مقادیر قابل انتخاب است طول این توصیف گر بسته به نیاز قابل تنظیم است و این آزادی عمل بالایی برای استفاده در کاربردهای مختلف به آن می بخشد. با آزمایش بر یک مولکول بیست و چهار اتمی که زوایای ساده و دوسطحی قابل تغییری دارد، نشان دادیم که تغییر در هندسه مولکول توسط مؤلفه های بردار توصیف گر اتم های درگیر به صورت مؤثری دیده می شوند.

نمونه ای از کاربرد این ابزار محاسباتی در پیش بینی سهم های اتمی از بار الکتریکی یا انرژی و مثل آنها در یک سامانه بس اتمی است. این طور پیش بینی ها اگر دقت کافی داشته باشند می توانند جای محاسبه مستقیم با روش های پرهزینه کلاسیک را بگیرند که این امر معمولاً با استفاده از تکنیک های یادگیری ماشین عملی می شود. در این مقاله یک مثال از این کاربردها را برای تخمین بارهای اتمی در ترکیب و تحول آن تحت اثر تغییرات ساختار هندسی ارائه دادیم. موفقیت قابل توجه و دقت بالای به دست آمده در این مثال به طور حیاتی مدیون عملکرد صحیح توصیف گر کولنی است. این توصیف گر برای کاربرد در سامانه های تناوبی و بلورین به راحتی قابل تعمیم است.

توضیحات

برنامه های کامپیوتری برای محاسبه توصیف گر و آزمایش یادگیری ماشین به روش KRR به زبان فرترن نوشته شده اند و به همراه راهنمای استفاده و داده های نمونه مربوط به شکل ها برای استفاده در مقاصد

- [14] م. مسجدی، ز. محمدی، م. اعلائی، ا. عبدالحسینی سارسری، ت. نیهاس، اثر اندازه بر طیف اپتیکی نانومولکول‌های رنگدانه‌ای کومارین با استفاده از رهیافت‌های تابعی چگالی تنگابست وابسته به‌زمان و تابعی چگالی وابسته به‌زمان فوق سریع، پژوهش سیستم‌های بس‌ذره‌ای **6** (1395) 15-31.
doi.org/10.22055/jrmb.2016.12449
- [14] I. Abdolhosseini Sarsari, T. Niehaus, Size effect on optical spectrum of coumarin nano molecule dyes via TD-DFTB and turbo-TDDFT approaches, *Journal of Research on Many-body Systems* **6** (2016) 5-31.
doi.org/10.22055/jrmb.2016.12449
- [15] V.Q. Vuong, J. Akkarapattiakal Kuriappan, M. Kubillus, J.J. Kranz, T. Mast, T.A. Niehaus, S. Irle, M. Elstner: Parametrization and benchmark of long-range corrected DFTB2 for organic molecules, *Journal of Chemical Theory and Computation* **14** (2017) 115-125.
doi.org/10.1021/acs.jctc.7b00947
- [10] F. Pietrucci, W. Andreoni, Graph theory meets ab initio molecular dynamics: atomic structures and transformations at the nanoscale, *Physical review letters* **107**, (2011) 085504.
doi.org/10.1103/PhysRevLett.107.085504
- [11] M. Rupp, A. Tkatchenko, K.R. Müller, O.A.V. Lilienfeld, Fast and accurate modeling of molecular atomization energies with machine learning, *Physical review letters* **108**, (2012): 058301.
doi.org/10.1103/PhysRevLett.108.058301
- [12] K. Vu, J.C. Snyder, L. Li, M. Rupp, B.F. Chen, T. Khelif, K.R. Müller, K. Burke, Understanding kernel ridge regression: Common behaviors from simple functions to density functional, *International Journal of Quantum Chemistry* **115** (2015) 1115-1128.
doi.org/10.1002/qua.24939
- [13] B. Aradi, B. Hourahine, Th. Frauenheim, DFTB+, a sparse matrix-based implementation of the DFTB method, *The Journal of Physical Chemistry A* **111** (2007) 5678.
doi.org/10.1021/jp070186p