ElectroLens: Understanding Atomistic Simulations through Spatially-Resolved Visualization of High-Dimensional Features

Xiangyun Lei\(^*\) Fred Hohman\(^*\) Duen Horng Chau\(^*\) Andrew J. Medford\(^*\)

Abstract
In recent years, machine learning (ML) has gained significant popularity in the field of chemical informatics and electronic structure theory. These techniques often require researchers to engineer abstract “features” that encode chemical concepts into a mathematical form compatible with the input to machine-learning models. However, there is no existing tool to connect these abstract features back to the actual chemical system, making it difficult to diagnose failures and to build intuition about the meaning of the features. We present ElectroLens, a new visualization tool for high-dimensional spatially-resolved features by a series of linked 3D views and 2D plots. The tool is able to connect different derived features and their corresponding regions in 3D via interactive selection. It is built to be scalable, and integrate with existing infrastructure.

Index Terms: Molecular Visualization—Visual Design—Coordinated and Multiple Views—Interaction Design

1 Introduction
Machine learning (ML) has seen lots of successful applications in the fields of chemical informatics and electronic structure theory in recent years [10][18]. Scientists have applied various fingerprinting approaches to describe atoms or their associated electronic environments in molecular systems, and trained ML models that connects these high-dimensional feature vectors to some properties of interest. One challenge that emerged from these studies is the lack of connection between these ML inputs and the underlying molecular systems. More specifically, with a set of feature vectors, there are no current tools for scientists to easily locate the corresponding atoms or electronic environments in the molecular systems. This is an issue for chemists, who typically develop strong “chemical intuition” through years of research and development experience. This intuition is based on the location and context of atoms or electrons in molecular systems, not on the derived mathematical features of ML models. The lack of connection prevents chemists from understanding the features intuitively. This makes it difficult for them to deduce the reason behind unsuccessful fingerprinting systems, or to improve them based on their chemical knowledge. Currently, researchers work around this mostly by trial and error: keep trying new fingerprinting system until the accuracy of the model
provides indirect evidence that it works. While some systematic approaches have emerged [15], they are still based on enumeration of many possible candidates and do not enable chemists to apply their domain knowledge to select or analyze the resulting fingerprints.

For example, a researcher might want to select features that differentiate the electronic structure of a bond between carbon (C) and nitrogen (N). This could be achieved by analyzing the electronic structure of a CH₂NO₂ molecule as shown in the left of Fig. 1 along with features derived using convolutional kernels [20] (right). By identifying and selecting sub-sets of these features the researcher can learn that when the combination of two features is within certain window (shown by the selection in the bottom right) the C=N bond is selected. This enables the researcher to identify this as a critical fingerprint in subsequent machine-learning models.

Visualization has strong potential to solve this problem, enabling users to not only look at the high-dimensional feature vectors, but also establish the connections between them and their corresponding regions in the actual system. The connection will help researchers link the features to their chemical domain knowledge for easier understanding. This is highlighted by the example above, where a researcher can use visualization to establish a connection between the numerical range of certain features and chemical concepts like C-N bonding, which can facilitate feature engineering and model improvement. Moreover, by considering the prediction accuracy of a ML model as an additional feature, the same tool can be used to directly visualize the performance of the model, identify problematic regions, and improve the model accordingly.

Contributions. In this study, we work with experts in ML models for electronic structure theory and contribute:

• ElectroLens, a 3D visualization tool for high-dimensional spatially-resolved features associated with atoms and electronic environments of molecular systems. The UI of ElectroLens, shown in Fig. 1, consists of two parts: the 3D view(s) (left) corresponding to Cartesian space, and 2D plots (right) corresponding to projections of the feature space.

• Interactive visualization design that displays high dimensional data through a series of 3D and 2D plots and connects them via interactive selection (Sect. 4). ElectroLens allows the user to make selections on the 2D plots, and the corresponding regions in real space will be highlighted in the 3D view.

• Scalable implementation that can process and visualize more than one million data points at a high frame rate of 60 FPS, on a commodity laptop computer.

• An open-source desktop application with Python bindings to the Atomic Simulation Environment (ASE) library [19] that is commonly used by scientists for managing atomistic simulations. ElectroLens is currently hosted on Github at medford-group.github.io/ElectroLens/ Tutorials and documentation are provided on the website.

• Usage scenarios illustrate how ElectroLens helped to decipher new descriptor systems and ML model performance in our research on ML models for electronic and atomistic structures.

2 BACKGROUND

Fingerprinting approaches are used to construct descriptive features to capture different aspects of molecular systems in the communities of ML electronic structure theory and chemical informatics. One common approach is to fingerprint molecular structures, leading to atom-centered features that capture the local chemical environment of a specific atom. There are numerous schemes including overlap integrals of Gaussians or atomic orbitals [4,8], Zernike polynomials [17], and others [11]. The result of these schemes is a high (typically >10) dimensional vector that describes each atom, which can be described as vectors corresponding to points on an irregular grid. An alternative approach is to examine the electronic structure surrounding the atoms instead of the atoms themselves. One strategy to achieve this is to treat the electron cloud as a voxelized 3D image, and to construct features to describe each voxel. For example, the recently-developed MCSH descriptors work in this way [20]. In this case, the resulting feature vectors correspond to points on a regular grid. In both cases the problem involves high-dimensional features that correspond to points in 3D Cartesian space.

The existing fingerprinting systems have been the foundation of numerous successful ML models for predicting the energies from atomic configurations [7,10], or energy contributions from local electronic structure [20]. However, there is a significant challenge in understanding the physical or chemical meaning of the features which are typically based on mathematical transformations rather than physical derivations. This problem will likely increase with the rise of deep learning techniques where the features are determined by the algorithm, making it even harder to assign specific meaning. Visualization provides a promising route to gain intuition about big, high-dimensional data sets and corresponding ML models [13,22]. While there are numerous software packages available for visualizing atomistic data sets [2,3,9,14,23–25,28], they have all been developed with chemical properties in mind, and are optimized for visualizing chemical concepts such as atom types, chemical bonds, and electron density. However, none are optimized to visualize the high-dimensional features used in ML models. Most programs are limited to visualizing a maximum of 2 features other than position (typically element type and radius) with a single 3D view, and often become sluggish when visualizing more than ten thousand data points. The data sizes required for ML models typically exceed this limit in terms of number of data points and dimensions per point.

3 DESIGN CHALLENGE

To establish an intuitive link between feature vectors and actual systems, we have worked with domain experts to identify six design challenges that are currently not addressed by other tools in the electronic structure theory or chemical informatics communities.

C1 Visualizing high-dimensional feature vectors. The typical fingerprinting systems used in research generates tens or even hundreds of features per atom or grid point. Visualization of the distribution and correlations between these features is crucial to the understanding of their meaning.

C2 Connecting features and corresponding Cartesian space. It is critical to visualize the features in their corresponding chemical context. A seamless connection between these alternative representations of the system is important to grasp the meanings of the features. Unfortunately, there is no existing tool in the community that establishes these connections visually.

C3 Comparing features across different systems. Assessing the generality of a given relationship between a feature vector and a chemical system requires comparison of multiple systems simultaneously. For example, if a feature is found to correspond to the C=O bonding region in a CO₂ molecule, it would be of interest to know if the same feature also corresponds to the C=O bond in a similar molecule such as HCOOH. However, most existing visualization tools for atomistic electronic structure focus on visualizing a single system at a time.

C4 Visualizing atomistic and electronic structures simultaneously. The atomic structure and electronic structure are intimately linked, but are represented with different data structures: irregular grids for atoms, and regular grids for electronic environments. The ability to simultaneously visualize the two data types, along with features that describe environments within these two related structures, will provide a new route to building intuitive connections between the two representations.

C5 Handling large datasets. Datasets used for ML training are often very large, containing millions of data points with tens of dimensions or more. Most tools for visualizing atomistic and electronic structure data do not scale well to datasets of extreme
WebGL-based technology is lightweight and enables high frame-rate (60fps) rendering of millions of points with a GPU [C5]. It is wrapped into a standalone desktop app that supports Windows, MacOS and Linux with the Electron JS library. The bindings provide an alternate strategy for navigating the high-dimensional space. They utilize principal component analysis (PCA) reduce the dimensionality of the features and plot the resulting projection using a scatter plot. These 2D plots can be added to the right-hand view on the fly with common transformations like “log10” for better visualization. Users can also choose to simultaneously plot features corresponding to atomic and electronic environments to assess connections between the two [C4]. One key functionality of the scatter plots and dimension-reduction plots is the ability to select regions in the 2D plots and see how these features are localized in 3D Cartesian space [C2], as highlighted in Fig. 1. ElectroLens supports exporting these sub-sets of points for further analysis, as well as saving specific features without loss of information [C1], a series of connected 2D plots are used. Three kinds of 2D plots are implemented: scatter plots, correlation plots and dimension-reduction plots. The scatter plots are heat maps where two features of choice are plotted and the color of the heat map corresponds to the amount of points at a given region of feature space. The correlation plot is a visual representation of the correlation matrix of the features. This provides a fast way to explore the high-dimensional feature space and identify interesting combinations, since feature combinations with low correlation tend to contain the most information. The dimension-reduction plots provide an alternate strategy for navigating the high-dimensional space. Users can principal component analysis (PCA) to reduce the dimensionality of the features and visualize different views of the results. Therefore, we have also created Python bindings for ElectroLens by wrapping with the CEFPythy library [C6]. The bindings are designed to follow the structure of the ASE Python library [L9], which is widely used by researchers in the field and contains classes corresponding to common data structures like atom positions [C6]. ASE also contains numerous translators for reading/writing a range of different file types into standardized data structures. By leveraging Python and ASE, ElectroLens is compatible with a wide range of file formats, and easy to use for anyone familiar with ASE.

4.2 Case Studies

This section presents two example applications of ElectroLens, presenting how it facilitated and accelerated research in ML and chemistry. They are: (1) Diagnosing the failure of a neural network for chemistry. They are: (1) Diagnosing the failure of a neural network for chemistry. They are: (1) Diagnosing the failure of a neural network for chemistry. They are: (1) Diagnosing the failure of a neural network for chemistry. They are: (1) Diagnosing the failure of a neural network for chemistry. They are: (1) Diagnosing the failure of a neural network for chemistry. They are: (1) Diagnosing the failure of a neural network for chemistry. They are: (1) Diagnosing the failure of a neural network for chemistry. They are: (1) Diagnosing the failure of a neural network for chemistry.
4.2 Constructing ML models from electron density

ElectroLens was originally developed to support a research project to establish a ML model to predict the exchange-correlation energy, a key quantity needed in density functional theory (DFT) [29]. Briefly, DFT is a widely-used technique for simulating the electronic structure of molecular systems. The formalism is based on a powerful theorem proving a one-to-one mapping between electron density and ground-state energy [30]. However, the connection to one part of the energy, known as the exchange-correlation (xc) energy, is unknown. Our work focused on using neural network (NN) to empirically learn this connection. As a first step, a NN model was trained between the local electron density at a point and the corresponding xc energy at the same point [27]. However, the resulting accuracy was insufficient, and ElectroLens was used to diagnose this result.

**View customization and double encoding.** ElectroLens supports encoding features or properties such as the force magnitude in the 3D visualization frame. This enables views like the one shown in Fig. 4 where atoms with excessively large forces can be easily identified, as they are colored by the extreme ends of the chosen color map. This can be used to construct new training data that contains atomic environments similar to those with large errors.

**Interactive selection with atomistic features.** In addition to visualizing errors in the 3D frame, it is also possible to plot them in the 2D frames, along with other input features to the model. This provides a route to identify regions of the feature space that have not been sampled, facilitating the generation of new training data or implementation of methods to raise errors when the model moves outside the domain of the training data.

The main lesson learned from this study is that the use of double-encoding features using color or size can assist researchers in identifying interesting patterns. This is not standard in visualizing chemical data, since size and color are typically determined by the element.

5 Conclusion and Future Work

A new visualization tool called ElectroLens has been developed to analyze high-dimensional features derived from 3D datasets. The tool has been developed in the context of atomic and electronic structure data, corresponding to two different representations (irregular and regular grids). ElectroLens is built to be efficient with large datasets easy to use and integrate with existing infrastructure for atomistic simulations. ElectroLens was applied and tested in multiple scenarios, leading to improved ML models for exchange-correlation energies and atomistic force fields (Sect. 4.2). Future work includes implementing other projections into ElectroLens, as well as studying methods to automatically infer informative feature combinations. We also plan to focus on applications of the tool to the increasingly popular NN force-fields for molecular dynamics simulations [6,10,18,21], including improved handling of time-dependent data sets. Further, we expect that ElectroLens may be useful for a wider set of problems in chemistry, physics, and engineering involving spatially-resolved high-dimensional data.

Acknowledgments

This work is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences Computational Chemical Sciences program under Award Number DE-SC0019410, with
support for atomistic feature visualization and integration with the AMPTorch code supported by Award Number DE-SC0019441. The authors are also grateful for a GPU generously provided by the NVIDIA Corporation through the GPU Grant program.

REFERENCES

[1] Electron. https://electronjs.org
[2] Jmol: an open-source java viewer for chemical structures in 3d. http://www.jmol.org/
[3] Tomviz. https://tomviz.org/
[4] A. P. Bartók, R. Kondor, and G. Csányi. On representing chemical environments. Phys. Rev. B, 87:184115, May 2013.
[5] A. D. Becke. Perspective: Fifty years of density-functional theory in chemical physics. The Journal of Chemical Physics, 140(18):18A301, 2014.
[6] J. Behler. Constructing high-dimensional neural network potentials: A tutorial review. International Journal of Quantum Chemistry, 115(16):1032–1050, 2015.
[7] J. Behler. Perspective: Machine learning potentials for atomistic simulations. The Journal of Chemical Physics, 145(17):170901, 2016.
[8] J. Behler and M. Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. Phys. Rev. Lett., 98:146401, Apr 2007.
[9] M. Biasini, T. Schmidt, S. Bienert, V. Mariani, G. Studer, J. Haas, N. Johner, A. D. Schenk, A. Philippson, and T. Schwede. OpenStructure: an integrated software framework for computational structural biology. Acta Crystallographica Section D, 69(5):701–709, May 2013.
[10] V. Botu, R. Batra, J. Chapman, and R. Ramprasad. Machine learning force fields: Construction, validation, and outlook. The Journal of Physical Chemistry C, 121(1):511–522, 2017.
[11] A. Chandrasekaran, D. Kamal, R. Batra, C. Kim, L. Chen, and R. Ramprasad. Solving the electronic structure problem with machine learning. npj Computational Materials, 5:22, February 2019.
[12] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. Phys. Rev., 136:B864–B871, Nov 1964.
[13] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau. Visual analytics in deep learning: An interrogative survey for the next frontiers. CoRR, abs/1801.06889, 2018.
[14] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. Journal of Molecular Graphics, 14:33–38, 1996.
[15] G. Imbalzano, A. Anelli, D. Giofre, S. Klees, J. Behler, and M. Ceriotti. Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. The Journal of Chemical Physics, 148(24):241730, Jun 2018.
[16] R. O. Jones. Density functional theory: Its origins, rise to prominence, and future. Rev. Mod. Phys., 87:897–923, Aug 2015.
[17] A. Khorshidi and A. A. Peterson. Amp: A modular approach to machine learning in atomistic simulations. Computer Physics Communications, 207:310 – 324, 2016.
[18] J. R. Kitchin. Machine learning in catalysis. Nature Catalysis, 1:230, Apr 2018.
[19] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christiansen, M. Dukh, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Marsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schitt, M. Strange, K. S. Thysgesen, T. Vegge, L. Vihelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen. The atomic simulation environment—a python library for working with atoms. Journal of Physics: Condensed Matter, 29(27):273002, 2017.
[20] X. Lei and A. J. Medford. Design and analysis of machine learning exchange-correlation functionals via rotationally invariant convolutional descriptors. Phys. Rev. Materials, 3:063801, Jun 2019.
[21] Y. Li, H. Li, F. C. Pickard, B. Narayanan, F. G. Sen, M. K. Y. Chan, S. K. S. Sankaranarayanan, B. R. Brooks, and B. Roux. Machine learning force field parameters from ab initio data. Journal of Chemical Theory and Computation, 13(9):4492–4503, 2017. PMID: 28800233.
[22] S. Liu, D. Maljovec, B. Wang, P. Bremer, and V. Pascucci. Visualizing high-dimensional data: Advances in the past decade. IEEE Trans-