From Shape to Structure: VAEs Decipher Stable Molecules from Hamiltonians

Nate Simmons
ECS Department
University of Texas at Dallas
Richardson, Texas, USA
nis190000@utdallas.edu

Rahul Kolla
ECS Department
University of Texas at Dallas
Richardson, Texas, USA
rxk190056@utdallas.edu

Sanjith Chockan
ECS Department
University of Texas at Dallas
Richardson, Texas, USA
sxc180101@utdallas.edu

James Dao
ECS Department
University of Texas at Dallas
Richardson, Texas, USA
jkd190003@utdallas.edu

Krishan Bansal
ECS Department
University of Texas at Dallas
Richardson, Texas, USA
kcb160130@utdallas.edu

Thomas Drablos
ECS Department
University of Texas at Dallas
Richardson, Texas, USA
tsd130130@utdallas.edu

***Abstract*—In this paper we present a Variational Autoencoder (VAE) framework for analyzing the structure of molecules and their corresponding electronic properties. The VAE utilizes a loss function called Evidence Lower Bound (ELBO). It will incorporate the mean squared error (MSE) and Kullback-Leibler (KL) divergence to ensure informative representations. The dataset chosen will include molecules characterized by electronic properties, which are derived from density functional theories (DFT). Feeding this dataset into the model will allow for molecular conformation analysis and its characteristics.**

***Keywords*—Molecular Conformation, Training, Algorithm, Variational Autoencoder, Kullback-Leibler, Hamiltonian**

I. Introduction

A. Background

In the modern era, the synthesis of drugs has become intimately related with fundamental concepts in quantum chemistry. Developing drugs which have specific biological effects requires one to know the underlying quantum properties of the bonds within a molecule of interest. This is because the way a molecule’s electronic system is arranged dictates important biochemical properties such as equilibrium and membrane transport. However, oftentimes researchers do not have all the information needed to produce specific molecules. When creating new drugs, the protein which the drug should target is often known. This information in turn gives an idea of what atoms/functional groups would be best for targeting the protein of interest.

The issue is that despite knowing the most likely atomic content and possible atomic arrangements of a novel drug, it is not clear what the best electronic arrangement of the atoms are. Knowing the spatial arrangement of atoms is not enough to predict the chemical properties of a molecule. Instead, scientists predict these properties via calculations based on the electronic structure of a molecule. For a given molecule there are generally hundreds of possible conformations (arrangements) of atoms, and each arrangement has a different energy level based on its electronic structure. Computing the energy for different conformations of a molecule is the key to creating stable compounds. Analytically solving the many particle Schrodinger equation is often computationally intractable, and so is using traditional density function theory approximations [1].

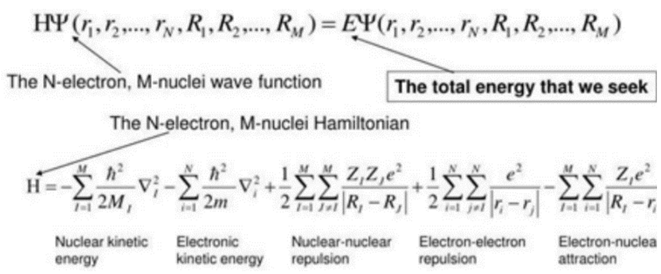
B. Problem Definition

Thus, the problem of finding the energy information for a given molecule lends itself to a machine learning based approach which relies on learning underlying patterns in the energy (represented as a Hamiltonian matrix) of molecular systems. Being able to accurately reconstruct Hamiltonian matrices is the first step in finding the energy of different atomic arrangements, which in turn allows researchers to design drugs which meet their specific requirements [2].

II. Theoretical and Conceptual Study

A. Techniques and Algorithms

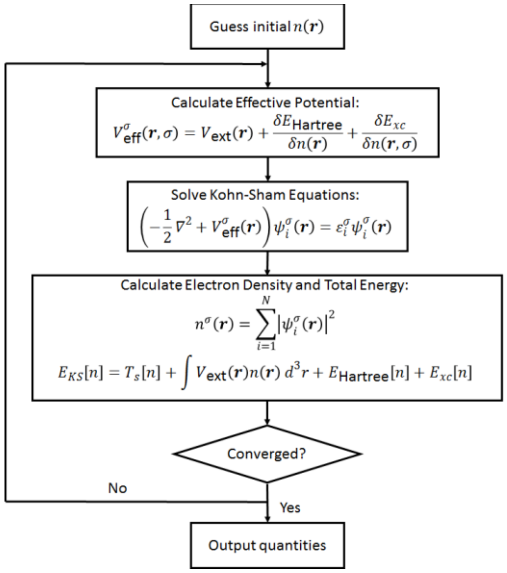
When solving the Schrodinger equation for a system, we get the possible quantum states for a given arrangement of atoms and the energy of these states. We require the Hamiltonian operator (total energy) of the system for this operation, and when we solve the Schrodinger equation using the Hamiltonian we get a ranking of each set of quantum states based on their energy level. In the case of quantum chemistry, we consider the quantum states as the positions of electrons within an atom. These quantum states provide much more information about the energy of a system rather than simply using the atomic content along with their spatial arrangement. While the positions of atoms around the nucleus for conformation are fixed, the possible quantum states of electrons (i.e. their spatial distributions) are not, and certain distributions are more stable/likely than others. The Hamiltonian for a many particle system describes the total potential and kinetic energy of the system based on the atoms within it, and the spatial arrangement of these atoms (conformation). Solving the time-independent Schrodinger equation we get the wave functions (eigenstates) of the various electrons within this atom, which describe the probability amplitude of finding those electrons at a certain point in space. We use the eigenvalues of the different wavefunctions to get the total energy of the system [3]. The analytical many particle Schrodinger equation for N electrons and M nuclei is shown below in figure 1.



$H\Psi(r_1, r_2, \ldots, r_N, R_1, R_2, \ldots, R_M) = E\Psi(r_1, r_2, \ldots, r_N, R_1, R_2, \ldots, R_M)$
The N-electron, M-nuclei wave function
The total energy that we seek
The N-electron, M-nuclei Hamiltonian
$H = -\sum_{I=1}^{M}\frac{\hbar^2}{2M_I}\nabla_I^2 - \sum_{i=1}^{N}\frac{\hbar^2}{2m}\nabla_i^2 + \frac{1}{2}\sum_{I=1}^{M}\sum_{J\neq I}^{M}\frac{Z_I Z_J e^2}{|R_I - R_J|} + \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{e^2}{|r_i - r_j|} - \sum_{I=1}^{M}\sum_{i=1}^{N}\frac{Z_I e^2}{|R_I - r_i|}$
Nuclear kinetic energy
Electronic kinetic energy
Nuclear-nuclear repulsion
Electron-electron repulsion
Electron-nuclea attraction

Fig. 1 Many particle Schrodinger equation [4]

The issue with this analytical approach is that it is impossible for systems that are more complex than a few electrons, and thus numerical methods which approximate the wavefunctions of many particle systems have been implemented since the time of Schrodinger. These methods fall under the general quantum computation category of Density Functional Theory (DFT). However, even these numerical methods are still quite slow, especially for complex systems. In this study, we utilize Hamiltonian matrices computed via the output of the Kohn-Sham method as training data and attempt to reconstruct them via sampling from a latent space. The Kohn-Sham method approximates single electron orbitals/wavefunctions and the potential energies that each electron in the system experiences, from this data we can then create a diagonalized Hamiltonian matrix which is used to calculate the energy values of the orbitals [1]. This general process is shown in the diagram below).



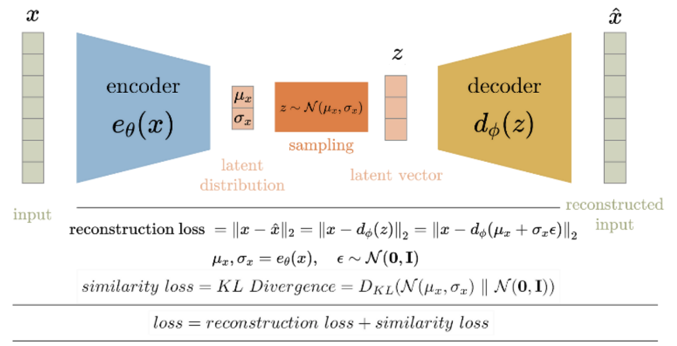
Guess initial $n(\mathbf{r})$
Calculate Effective Potential:
$V_{\text{eff}}^{\sigma}(\mathbf{r},\sigma) = V_{\text{ext}}(\mathbf{r}) + \frac{\delta E_{\text{Hartree}}}{\delta n(\mathbf{r})} + \frac{\delta E_{xc}}{\delta n(\mathbf{r},\sigma)}$
Solve Kohn-Sham Equations:
$\left(-\frac{1}{2}\nabla^2 + V_{\text{eff}}^{\sigma}(\mathbf{r})\right)\psi_i^{\sigma}(\mathbf{r}) = \varepsilon_i^{\sigma}\psi_i^{\sigma}(\mathbf{r})$
Calculate Electron Density and Total Energy:
$n^{\sigma}(\mathbf{r}) = \sum_{i=1}^{N}|\psi_i^{\sigma}(\mathbf{r})|^2$
$E_{KS}[n] = T_s[n] + \int V_{\text{ext}}(\mathbf{r})n(\mathbf{r})\,d^3r + E_{\text{Hartree}}[n] + E_{xc}[n]$
Converged?
No
Yes
Output quantities

Fig. 2 Kohn-Sham method [5]

The KS method was used to compute Hamiltonian matrices for the training dataset. A Hamiltonian matrix was computed for each possible conformation of atoms of a given molecule. The energy of the system was calculated using the Psi4 quantum chemistry library [4].

B. Approach

Our approach to reconstructing the Hamiltonian Matrices of many particle systems relies on a variational autoencoder neural network (VAE) trained on the nablaDFT dataset. Variational autoencoders are generative models which rely on an encoder to map information into a probabilistic latent space (defined by a mean and variance), and a decoder to sample from the latent space to reconstruct the input data. This approach has advantages not only due to the probabilistic nature of quantum mechanics, but also because the latent space has less dimensionality than the many features which are normally extracted from Hamiltonian matrices via deep neural networks. The loss function of the VAE consists of 2 parts: one which measures the reconstruction loss between input samples and the output reconstructed by the decoder (reconstruction loss), and one which fits the latent space distribution to the standard Gaussian normal distribution N(0,1) [4]. The loss and general VAE architecture are shown below.



x
$\hat{x}$
encoder
$e_\theta(x)$
μ_x
σ_x
$z \sim \mathcal{N}(\mu_x, \sigma_x)$
z
decoder
$d_\phi(z)$
sampling
latent distribution
latent vector
input
reconstructed input
reconstruction loss $= \|x - \hat{x}\|_2 = \|x - d_\phi(z)\|_2 = \|x - d_\phi(\mu_x + \sigma_x\epsilon)\|_2$
$\mu_x, \sigma_x = e_\theta(x), \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$similarity\ loss = KL\ Divergence = D_{KL}(\mathcal{N}(\mu_x, \sigma_x) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I}))$
$loss = reconstruction\ loss + similarity\ loss$

$$
\begin{aligned}
(g^*, h^*) &= \underset{(g,h)\in G\times H}{\arg\min}\ KL(q_x(z), p(z|x)) \\
&= \underset{(g,h)\in G\times H}{\arg\min}\left(\mathbb{E}_{z\sim q_x}(\log q_x(z)) - \mathbb{E}_{z\sim q_x}\left(\log\frac{p(x|z)p(z)}{p(x)}\right)\right) \\
&= \underset{(g,h)\in G\times H}{\arg\min}\left(\mathbb{E}_{z\sim q_x}(\log q_x(z)) - \mathbb{E}_{z\sim q_x}(\log p(z)) - \mathbb{E}_{z\sim q_x}(\log p(x|z)) + \mathbb{E}_{z\sim q_x}(\log p(x))\right) \\
&= \underset{(g,h)\in G\times H}{\arg\max}\left(\mathbb{E}_{z\sim q_x}(\log p(x|z)) - KL(q_x(z), p(z))\right) \\
&= \underset{(g,h)\in G\times H}{\arg\max}\left(\mathbb{E}_{z\sim q_x}\left(-\frac{||x - f(z)||^2}{2c}\right) - KL(q_x(z), p(z))\right)
\end{aligned}
$$

Fig. 3 Loss and VAE architecture [6]

The mean and variance of the VAE's latent must be parameterized in the manner shown below to be differentiable and thus compatible with backpropagation.

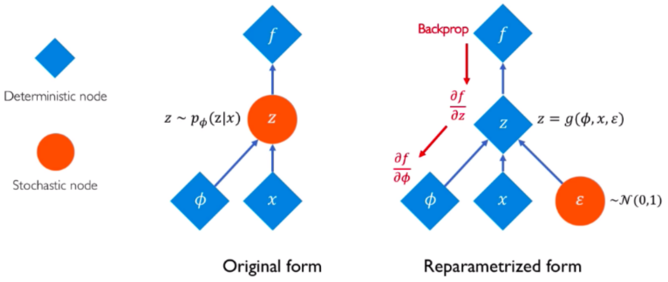


Fig. 4 Parameterized mean and variance of VAE's latent [7]

We take in the Hamiltonian matrices as training samples and pad them to a standard size of 415x415. The matrices are then normalized to maintain low weight values (gradient clipping was also applied). For computational efficiency, we flatten the Hamiltonian matrices into a nx1 vector rather than use convolution layers (we strongly recommend applying convolution layers for the clinic use of this network). We update the weights of the network via traditional stochastic gradient descent, though ADAM would work just as well if not even better. The entire network was written using numpy (including the backpropagation) as an intellectual exercise, however for a real application of this research Pytorch/Keras libraries would be ideal. We train the network with batches of 32 matrices, from a total sample set of 1000 matrices. The entire nablaDFT dataset contains millions of conformations, totaling 7 terabytes, so for the purposes of this project we select a 6Gb subset.

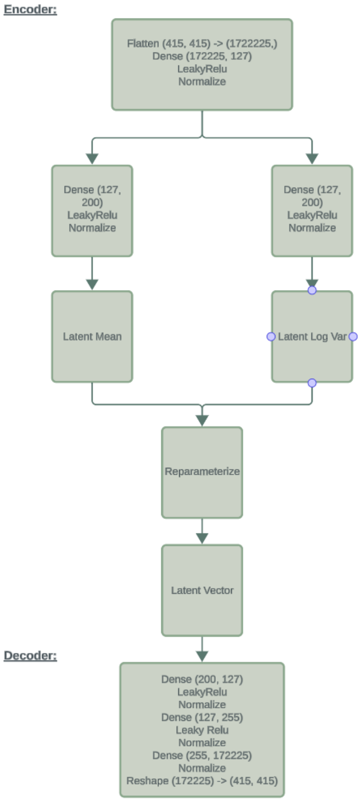
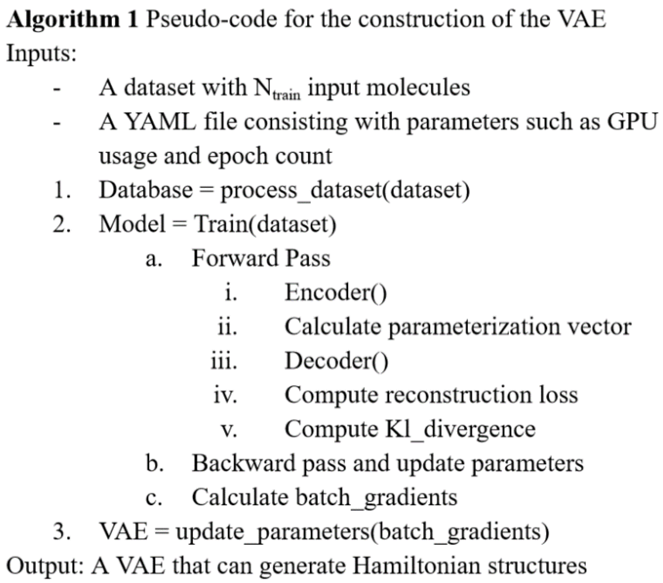


Fig. 5 Layout of VAE encoder and decoder

Above is the outline of the encoder and decoder. Below is the algorithm to construct a VAE utilizing the above architecture:



Algorithm 1 Pseudo-code for the construction of the VAE

```
Inputs:
    -  A dataset with $N_{train}$ input molecules
    -  A YAML file consisting with parameters such as GPU
       usage and epoch count
    1. Database = process_dataset(dataset)
    2. Model = Train(dataset)
        a. Forward Pass
            i.   Encoder()
            ii.  Calculate parameterization vector
            iii. Decoder()
            iv.  Compute reconstruction loss
            v.   Compute Kl_divergence
        b. Backward pass and update parameters
        c. Calculate batch_gradients
    3. VAE = update_parameters(batch_gradients)
Output: A VAE that can generate Hamiltonian structures
```

Fig. 6 Algorithm overview

i. Update Parameters

Updates the weights for the Encoder and Decoder utilizing the Leaky ReLu Derivative which is computed as 1 if X > 0, otherwise the algorithm returns the alpha as defined by the Configurations file.

III. Results and Analysis

A. Results

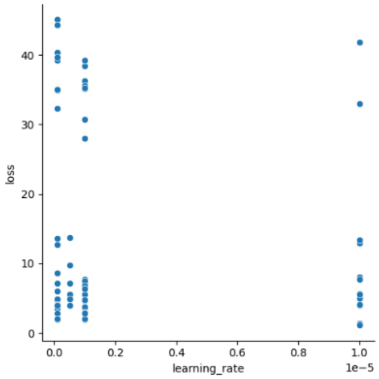


Fig. 7 Loss vs Learning Rate

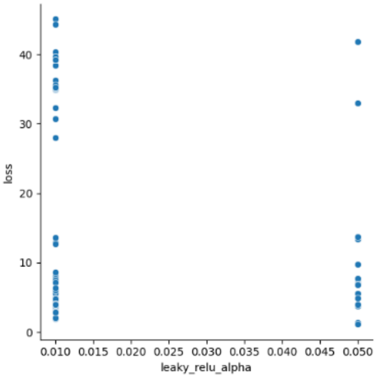


Fig. 8 Loss vs Leaky ReLu

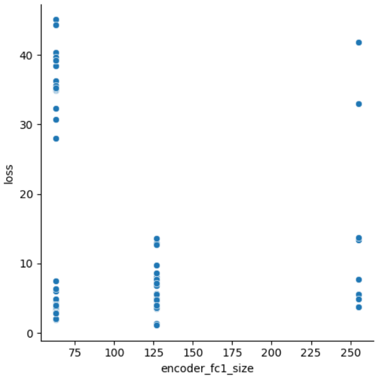


Fig. 9 Loss vs Encoder

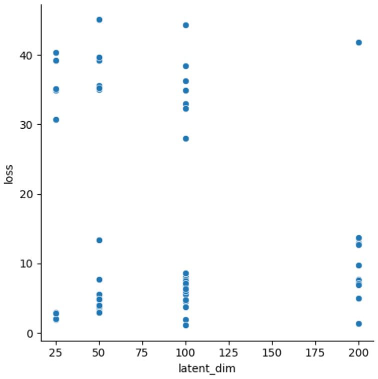


Fig. 10 Loss vs Latent Dimension

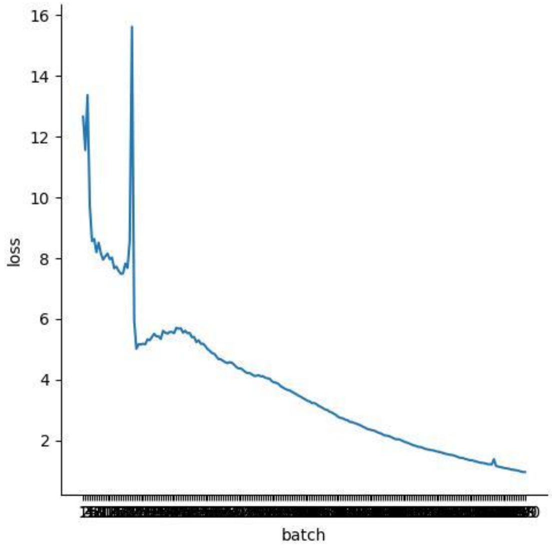


Fig. 11 Loss vs Batch

B. *Analysis*

The figures above detail the analysis of the VAE performance, based on our metric of loss, under various hyperparameter configurations. For the latent dimension graph, there doesn't seem to be a correlation between the loss and latent dimension size as it increases. The variation in the loss is less in higher dimensions, but there is still not a strong correlation. The encoder and decoder graphs are redundant and therefore only one graph is included. Each is composed of similar patterns where lower losses correlate to larger layer sizes. One noticeable difference for the encoder would apply to the range of sizes. The loss can be seen to have a significant variation from the decoders tested. The Leaky Relu is another hyperparameter that was tested. Increasing the alpha value from the activation function produced lower losses. We can perceive that increasing alpha would be advantageous in reducing loss. Next, we observe the learning rate hyperparameter. There is a clear relationship between lower loss and low learning rates. However, there is a spike in loss at around 1e-5. One issue we faced was exploding gradients. This was eventually addressed by normalization layers after each activation function within encoder and decoder. Normalizing the entire dataset also facilitated faster convergence.

IV. Conclusion

This work explored the application of a VAE framework to analysis of the relationship between molecular conformations and their electronic properties. The VAE leveraged ELBO loss function as a metric, and incorporated reconstruction accuracy and latent representations. There was no clear correlation between latent dimension and loss, and the larger layer sizes in the encoder and decoder led to lower losses. Leaky ReLU activation function with a high alpha value and low learning rates generally yielded better performance. Issues such as exploding gradients were addressed through normalization layers. This summarizes the key findings of the work.

V. FUTURE WORK

The current VAE framework offers a promising foundation for analyzing molecular conformations and their electronic properties. However, there are several areas for future work. Firstly, convolutional neural networks (CNNs) have made remarkable strides in the field of capturing image data relationships [5]. Here, the high dimensionality of electronic properties may also be effectively modeled by such networks. Therefore, implementing convolutional layers into the existing VAE may lead to a more efficient and accurate representation of the molecular conformations. Next, future iterations of this work may incorporate Psi4[10]. Psi4 is an open-source quantum chemistry suite that permits accurate simulation of molecular properties [4]. This would enable the current VAE architecture to process the molecular structures and compute the electronic properties within its own training loop rather than accessing pre-computed data. Lastly, the VAE framework may eventually be extended to generate new molecule conformations. By sampling prior distributions, the VAE should theoretically be capable of discovering new drug candidates given relevant target properties.

REFERENCES

[1] Kuzma Khrabrov et al., “nablaDFT: Large-Scale Conformational Energy and Hamiltonian Prediction benchmark and dataset,” Physical chemistry chemical physics/PCCP. Physical chemistry chemical physics, vol. 24, no. 42, pp. 25853–25863, Jan. 2022, doi: https://doi.org/10.1039/d2cp03966d.

[2] P.-Y. Kao et al., “Exploring the Advantages of Quantum Generative Adversarial Networks in Generative Chemistry,” Journal of Chemical Information and Modeling, vol. 63, no. 11, p. 3307, Jun. 2023, doi: https://doi.org/10.1021/acs.jcim.3c00562.

[3] K. A. Baseden and J. W. Tye, “Introduction to Density Functional Theory: Calculations by Hand on the Helium Atom,” Journal of Chemical Education, vol. 91, no. 12, pp. 2116–2123, Nov. 2014, doi: https://doi.org/10.1021/ed5004788.

[4] eldora, “PPT - Quantum Mechanics in a Nutshell PowerPoint Presentation, free download - ID:5172062,” SlideServe, Oct. 05, 2014. https://www.slideserve.com/eldora/quantum-mechanics-in-a-nutshell#google_vignette (accessed May 08, 2024).

[5] A. Chen, “Figure 2,” ResearchGate. https://www.researchgate.net/figure/Iterative-computational-process-for-solving-the-Kohn-Sham-equation_fig11_341945597 (accessed May 07, 2024).

[6] A. Anwar, “Difference between AutoEncoder (AE) and Variational AutoEncoder (VAE),” Medium, Nov. 04, 2021. https://towardsdatascience.com/difference-between-autoencoder-ae-and-variational-autoencoder-vae-ed7be1c038f2

[7] S. Paul, “‘Reparameterization’ trick in Variational Autoencoders,” Medium, Apr. 06, 2020. https://towardsdatascience.com/reparameterization-trick-126062cfd3c3 (accessed May 08, 2024).

[8] F. Li, “FaustineLi/Variational-Autoencoders,” GitHub, Nov. 09, 2023. https://github.com/FaustineLi/Variational-Autoencoders (accessed May 07, 2024).

[9] L. Meneghetti, N. Demo, and Gianluigi Rozza, “A dimensionality reduction approach for convolutional neural networks,” Applied intelligence, vol. 53, no. 19, pp. 22818–22833, Jul. 2023, doi: https://doi.org/10.1007/s10489-023-04730-1.

[10] “Psi4 1.4: Open-Source Software for High-Throughput Quantum Chemistry”, D. G. A. Smith, L. A. Burns, A. C. Simmonett, R. M. Parrish, M. C. Schieber, R. Galvelis, P. Kraus, H. Kruse, R. Di Remigio, A. Alenaizan, A. M. James, S. Lehtola, J. P. Misiewicz, M. Scheurer, R. A. Shaw, J. B. Schriber, Y. Xie, Z. L. Glick, D. A. Sirianni, J. S. O’Brien, J. M. Waldrop, A. Kumar, E. G. Hohenstein, B. P. Pritchard, B. R. Brooks, H. F. Schaefer III, A. Yu. Sokolov, K. Patkowski, A. E. DePrince III, U. Bozkaya, R. A. King, F. A. Evangelista, J. M. Turney, T. D. Crawford, C. D. Sherrill, J. Chem. Phys. (2020). (doi: 10.1063/5.0006002).