ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

High-Throughput Screening of Promising Redox-Active Molecules with MolGAT

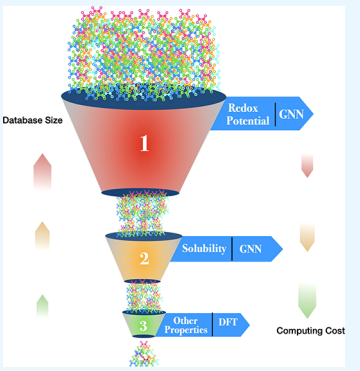
Mesfin Diro Chaka, Chernet Amente Geffe, Alex Rodriguez, Nicola Seriani, Qin Wu, and Yedilfana Setarge Mekonnen*

Cite This: *ACS Omega* 2023, 8, 24268−24278

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Redox flow batteries (RFBs) have emerged as a promising option for large-scale energy storage, owing to their high energy density, low cost, and environmental benefits. However, the identification of organic compounds with high redox activity, aqueous solubility, stability, and fast redox kinetics is a crucial and challenging step in developing an RFB technology. Density functional theory-based computational materials prediction and screening is a time-consuming and computationally expensive technique, yet it has a high success rate. To speed up the discovery of new materials with desired properties, machine-learning-based models can be trained on large data sets. Graph neural networks (GNNs) are particularly well-suited for non-Euclidean data and can model complex relationships, making them ideal for accelerating the discovery of novel materials. In this study, a GNN-based model called MolGAT was developed to predict the redox potential of organic molecules using molecular structures, atomic properties, and bond attributes. The model was trained on a data set of over 15,000 compounds with redox potentials ranging from −4.11 to 2.56. MolGAT outperformed other GNN variants, such as the Graph Attention Network, Graph Convolution Network, and AttentiveFP models. The trained model was used to screen a vast chemical data set comprising 581,014 molecules, namely OMDB, QM9, ZINC, CHEMBL, and DELANEY, and identified 23,467 potential redox-active compounds for use in redox flow batteries. Of those, 20,716 molecules were identified as potential catholytes with predicted redox potentials up to 2.87 V, while 2,751 molecules were deemed potential anolytes with predicted redox potentials as low as −2.88 V. This work demonstrates the capabilities of graph neural networks in condensed matter physics and materials science to screen promising redox-active species for further electronic structure calculations and experimental testing.



Downloaded via UNIVERSITA STUDI TRIESTE on October 9, 2023 at 09:20:40 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

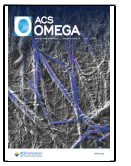
1. INTRODUCTION

Research into high-capacity, low-cost batteries has been sparked by the rapidly growing global demand for energy, with a focus on electric vehicles to reduce dependence on limited petroleum resources and advanced electrical energy storage devices needed for the electrical grid system to stabilize power supply fluctuations.[1] Electrochemical energy storage devices have gained attention due to their high energy density and low cost compared to conventional batteries over the past few decades. Although lithium ion batteries are the most popular and widely used type of energy storage device because of their excellent stability and storage capacity,[2] their widespread adoption has been impeded by the use of rare metals like cobalt, despite their popularity in various applications, particularly for portable electronic devices. Lithium and sodium ion batteries are limited by their high cost, slow charging, and low energy/power density when compared to gasoline.[3]

The field of rechargeable batteries is currently focused on developing next-generation batteries, such as metal−air batteries[4−6] and metal−sulfur batteries,[7] as well as enhancing the performance of existing batteries.[8,9] These batteries are mainly used in electric appliances including automobiles. However, redox flow batteries (RFBs) have emerged as a promising candidate for large-scale energy storage and electricity generation, owing to their modular and flexible design, low maintenance costs, extended cycling life, and eco-friendliness.[10−12] To achieve gigawatt-scale energy storage, RFB research has shifted toward utilizing low-cost redox species, such as transition-metal complexes, organic molecules,[13,14] and polymers.

RFBs are a type of rechargeable battery that consist of three main components: a stack of electrochemical cells, flow mechanisms, and energy storage tanks. At the surface of the RFB cells, which are made of materials such as graphite felt, carbon paper, or metal foam, redox reactions occur.[15] These reactions involve the electroactive materials that are dissolved

Received: February 26, 2023
Accepted: June 12, 2023
Published: June 30, 2023



ACS Publications

© 2023 The Authors. Published by American Chemical Society

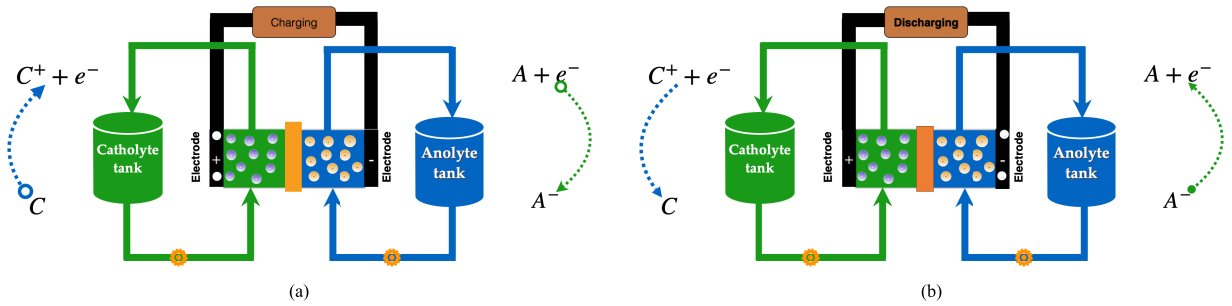


Figure 1. Schematic working principle of redox flow batteries (RFBs) (a) during charging and (b) during discharging

in supporting electrolytes and circulate between the tanks and corresponding compartments of the electrochemical cell. External pumps at the electrodes drive the flow of the electrolytes to achieve energy conversion between chemical and electrical energy.[10] The electrolytes are flowed back to the cell stacks to increase the battery's total state of charge. The volume of the two tanks where the anolyte and catholyte are stored represents the overall energy of the system, while the size of the stack cells separated by ion-exchange materials determines the system's power. Catholyte and anolyte refer to the redox-active substances dissolved in positive and negative electrolytes, respectively. During charging, the catholyte materials release electrons and undergo oxidation, while the anolyte materials receive electrons and undergo reduction.[15] On the other hand, ions diffuse through the membrane toward the opposite half-cell to balance the charge. The energy needed to charge RFBs is supplied by an external power source, such as solar PV installations. The discharge phase involves the reverse process, as illustrated in Figure 1.

The electrochemical behavior of organic materials in various environments is believed to aid in identifying substitute redox-active materials for RFBs that are more sustainable and perform better as RFBs have a higher energy density than conventional lead-acid batteries due to their acid-free operation. Finding environmentally friendly and effective substitute redox-active materials for RFBs can be challenging due to the intricate electrical combinations and many-body interactions in molecular structures, which make it difficult to represent them comprehensively.[16] The complexity of identifying materials with favorable redox potential, solubility, and chemical stability properties stems not only from the uncertainty of redox flow battery systems but also from the complexity of the information in chemical molecular systems, which is challenging to discern directly from chemical structures. Graph-based deep learning has gained attention in both theoretical and applied machine learning for scientific research, demonstrating improved performance in various scientific applications such as atomic reaction analysis, molecular property prediction, and molecular generation.[17]

The search for new materials is crucial to advancing technologies that are more cost-effective, useful, and sustainable. However, the vastness of the chemical space makes it impractical to search for new materials through exhaustive exploration.[18] To overcome this limitation, significant effort has been put into developing high-throughput ab initio simulations[19] that can calculate material characteristics. The development of next-generation RFBs with high cell voltage, energy density, cycle life, and power density requires the discovery of redox-active materials with high redox potential, high aqueous solubility, high stability, and faster redox kinetics.[13] Computational screening techniques have emerged as an effective alternative to trial and error experimentation. It is critical to discover novel electrolytes that perform efficiently to meet the growing interest in the development of improved electrochemical energy storage devices.[20] However, the conventional approach to creating energy materials is fraught with obstacles, including low success rates, prolonged time consumption, and exorbitant computational expenses. Consequently, screening of advanced materials and modeling their quantitative structure−activity relationships have recently become hot and trending topics in energy materials. Virtual screening is a computational technique that uses computer algorithms to identify compounds with high affinity for a specific target.[21] Although commonly used in drug discovery, virtual screening can be applied to any type of molecule, including electrolyte parameters that are universally significant for all types of batteries, such as redox potential, solubility, and stability. A typical screening method involves narrowing down a pool of candidates based on subsequent property assessments discovered through high-throughput density functional theory (DFT) or machine-learning calculations.

As obtaining an accurate absolute potential measurement is a difficult task, the redox potential values of both positive and negative electrodes are compared to that of hydrogen (H_2) which is established as a reference fixed at 0 V under standard conditions. Base metals are metals with a negative redox potential, while noble metals have a positive redox potential. Redox potential in aqueous solutions is an indication of the solution's inclination to acquire or relinquish electrons following the introduction of a new species. Typically, a solution with a lower (more negative) reduction potential will release electrons to the new species, whereas a solution with a higher (more positive) reduction potential will receive electrons from it. The thermodynamic basis of to predict the redox potential of redox-active species is the aqueous-phase redox reaction $M + 2H^+ + 2e^- \Leftrightarrow MH_2$, in which M is the redox-active molecular reactant species and MH_2 is the corresponding molecules that are produced as hydrogenated products as a result of their respective chemical reactions. The reaction energy ΔE_{rxn} of the redox couples has been calculated by

$$\Delta E_{rxn} = E(MH_2) - [E(M) + E(H_2)] \quad (1)$$

where $E(MH_2)$, $E(M)$, and $E(H_2)$ are the total energies of the reactants, the product molecules, and hydrogen molecules, respectively.

Through computational modeling, researchers can gain insights into the electrochemical characteristics of redox-active materials, leading to the identification of novel molecules with

https://doi.org/10.1021/acsomega.3c01295
ACS Omega 2023, 8, 24268−24278

enhanced properties. High-throughput computational screening enables the exploration of thousands of molecules to identify those with desirable properties, bypassing the need for costly and time-consuming experimental trial and error.[22] Machine learning proves to be a powerful tool for predicting molecular properties for high-throughput screening in novel materials discovery, as it can quickly analyze large data sets such as ChEMBL[23] and ZINC[24] and identify complex relationships between input features and target properties.[25−28] Compared to traditional DFT-based computational screening methods, machine learning-based high-throughput screening for novel materials discovery can offer significant advantages. DFT-based high-throughput screening methods are computationally expensive and time-consuming, whereas machine learning models can efficiently process large data sets without the need for expensive calculations.[29,30] Furthermore, machine learning models can continuously learn and improve through iterative training, providing more accurate predictions and better material discovery.

The application of machine learning in various fields has led to the development of models that can perform mentally taxing tasks and, in some cases, even surpass human performance,[31,32] as well as offer significant time and cost savings.[33] Among the various subsets of machine learning, deep learning involves training neural networks on large amounts of data to perform specific tasks, and models such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM) networks, gated recurrent units (GRUs), and graph neural networks (GNNs) are some of the major examples. Machine learning models have wide-ranging applications in condensed matter physics and materials science, including crystal structure prediction,[34] phase diagram determination,[35] aqueous solubility prediction,[36] fingerprint prediction,[37] and inverse design.[21,38−42]

Machine learning models that utilize descriptor-based features are models that use molecular descriptors as input features to make predictions about various molecular properties.[43] These descriptors are numerical values that represent different characteristics of a molecule such as its size, shape, electronic properties, and other physicochemical properties. They are derived from the molecular structure and are often calculated by using computational methods. The state-of-the-art molecular descriptor is extended-connectivity circular fingerprints (ECFP),[44] which apply a fixed hash function to the concatenated features of the neighborhood in the previous layer to provide the features for each layer. To treat these hashes as integer indices, a 1 is written to each node's fingerprint vector at the index indicated by the feature vector.[45] The Morgan algorithm[46] has been improved with circular fingerprints to make them invariant to atom-relabeling and encode the present substructures. Fingerprint representations focus on some characteristics of the chemical structure while ignoring others,[45] unlike models that rely on data to make decisions. A significant obstacle to widespread adoption is identifying appropriate model inputs, or "descriptors". These molecular descriptors (fingerprints) are often utilized as input in traditional quantitative structure−activity relationship (QSAR) property prediction, and then a specific deep learning architecture is used to train a model.[47] As a result of using structure-based descriptors, the models are limited by the same structural constraints as ab initio approaches when looking for new compounds.

Real-world data sets, including molecular structures, protein−protein interactions, brain connectome data, social networks, citation networks, and others, are frequently connected by graph structures. Consequently, it is crucial for machine learning research to extend deep neural networks to handle this type of input, but until recently, this topic has received little attention. In order to address this issue of lack of transferability in deep learning models for predicting chemical properties, graph-based models such as graph neural networks (GNN) have been developed. These models utilize molecular graphs as input to capture essential chemical information.[48] Several GNN variants[45,49−51] have been developed that generalize this process to irregular molecular graphs, a natural representation of chemical structures, similar to how convolutions are applied to regular data, such as text and images. Therefore, graph models, including graph convolutions, extract relevant features from graph structure descriptions, such as atom and bond characteristics and graph distances, to create molecular-level representations that can replace fingerprint descriptors in machine learning applications.[52]

Na et al.[53] developed a graph neural network with a graph feature vector-based attention mechanism to assess atomic significance. They combined the reverse graph self-attention (RGSA) method with machine learning-based methods for atomic significance estimation. Another interesting study, the attentive FP[54] model, uses a graph-based molecular representation that allows the model to learn nonlocal intramolecular interactions for specific prediction tasks. However, there have been few studies on the applicability of the GNN to predict electrolyte properties in RFBs. To address this gap, a molecular graph attention network (MolGAT) was proposed. The MolGAT model enhances the graph attention network (GAT) model by adding edge attributes of the molecular graph by using GNN modeling techniques. This model can predict the redox potential of organic electrolyte materials and it outperformed commonly used GNN variants in terms of MAE and RMSE errors. Finally, the trained model was utilized to screen promising novel redox-active species.

2. COMPUTATIONAL METHODS

2.1. Model Architecture. A pair (V, E) represents an undirected graph G, where V denotes the graph's nodes and E denotes the edges connecting them. A matrix A of size $|V|\times|V|$ represents the graph, with each element x_{ij} being either 1 or 0, indicating the presence or absence of an edge between nodes i and j. Since the graph is undirected, $x_{ij} = x_{ji}$ holds true for all elements of A. However, many graphs have additional information attached to their nodes and edges. In the case of a molecule represented as an undirected graph, the node label matrix encodes each node's atom type, while the edge label matrix encodes each edge's bond type. Hydrogen atoms are often excluded from the picture to simplify it as their placement can be inferred from fundamental chemical principles. Each node in the molecular graph corresponds to a chemical element (e.g., O, C, N, or H), and each edge represents the type of bond (single, double, triple, or aromatic) connecting the nodes. To encode these node and edge properties, a node feature matrix X of size $|V|\times f_V$ is defined for the node labels, where f_V is the length of the label vector of each node. Similarly, the edge feature matrix X_E of size $|E|\times f_E$ is defined for the edge labels, where f_E is the length of the label vector for each edge.

https://doi.org/10.1021/acsomega.3c01295
ACS Omega 2023, 8, 24268−24278

In molecular property prediction and atomic reaction analysis, GNNs have gained considerable attention due to their ability to utilize the relationships between data points in graph-structured data to generate an output. This approach has been shown to be more promising than conventional descriptor-based models, as demonstrated in studies such as refs 45 and 17. GNN eliminates the need for hand-crafted descriptors and/or fingerprints, particularly in graph-based problems, such as predicting chemical properties using graph-based representations of molecules.[55] The aim of GNN is to acquire a representation of each atom by transmitting messages recursively throughout the molecular graph, to compile information from connected bonds and neighboring atoms, and then to update the central atoms' states and carry out graph aggregation read-out operations. The message passing scheme is used to generalize the convolution operator over irregular domains. In this approach, the message moves from one node to the next when the dot product of the adjacency matrix is applied to the message vector (node feature matrix) and all nodes receive messages from their neighbors simultaneously. Therefore, each node embedding contains information about its k-hop neighborhood after k-iterations. The molecular graph data is usually presented in two formats: structural information concerning the representation of the molecular graph and atom and bond features from node-embedding and edge-embedding of molecular graphs.[56] The neural message passing approach produces a set of node embeddings that need to have a graph pooling aggregation for the embedding Z_G of the entire graph.[56] Finally, the representation can be used to predict the molecular properties.

Several GNN variants have been developed for various applications, such as Crystal Graph Convolutional Neural Network (CGCNN),[57] Neural Message Passing for Quantum Chemistry (MPNN),[58] MatErials Graph Network (MEGNet),[59] Modeling Rational Data with Graph Convolutional Networks (R-GCN),[60] and (EGNN).[61] While all of these are used in materials science and chemistry, they differ in their approaches and strengths. MPNN is well-suited for fixed-topology molecules, while MEGNet can handle varying graph structures and capture global information indirectly through its learned distance matrix. In contrast, CGCNN is designed for crystal structures and can extract features from both nodes and edges to predict properties such as formation energies and elastic moduli. R-GCN operates on directed graphs with only node features and uses a simplified version of spectral graph convolutions, while EGNN considers edge features by using a gated message passing scheme that involves passing messages along edges and updating node representations. The selection of MPNN, MEGNet, or CGCNN depends on the specific task and available data. MPNN and MEGNet are both suitable for predicting molecular properties, but MEGNet may be preferred for tasks where 3D structure is relevant. CGCNN is ideal for predicting properties of crystal structures, such as band gap and elastic constants. Although MPNN, MEGNet, and CGCNN are commonly used for material and crystal property prediction and molecular dynamics, R-GCN and EGNN have been used in recommendation systems, knowledge graphs, social network analysis, and 3D point cloud segmentation.

GCN[45] and GAT[62] are two widely used types of GNNs used in deep learning for extracting structural features from molecular graphs.[50,62] Both models utilize convolution layers to calculate new features based on the input features and graph structure but differ in their approach to neighborhood aggregation. GCN uses a nonparametric weight with a normalizing function, which limits its generalizability. In contrast, GAT captures node importance based on spatial topology by learning message-passing weights from hidden embeddings using attention scores, allowing for a data-driven approach to operator selection without prior assumptions. It is beneficial to use atomic properties and molecular structure as inputs in graph representation learning to allow the algorithm to determine relevant information. Although edges and nodes may have associated attributes, GCN and GAT mainly focus on node features and 1-dimensional edge features, disregarding n-dimensional edge features. GAT extends GCN with an attention mechanism that learns message-passing weights c_{ij} from hidden embeddings using attention scores $\alpha_{ij}^{(l)}$ in a data-driven approach, rather than fixing the operator a priori with a shared linear transformation weight matrix $\Theta \in \mathbb{R}^{F' \times F}$, given by the following formula:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i) \cup \{i\}} \alpha_{ij}^{(l)} \Theta^{(l)} h_j^{(l)} \right) \tag{2}$$

This "self-attention mechanism (weighting factor)" α_{ij} is based on a simple principle that atoms (nodes) should not all have the same relevance. It is computed as the softmax normalized inner product between a learnable weight vector W_a and the concatenation of the transformed hidden embedding of the two nodes written as

$$\begin{aligned} \alpha_{ij}^{(l)} &= \mathrm{softmax}(a_{ij}^{(l)}) \\ &= \frac{\exp(a_{ij}^{(l)})}{\sum_{k \in \mathcal{N}(i)} \exp(a_{ik}^{(l)})} \\ &= \frac{\exp(\mathrm{LeakyReLU}(W_a^T \cdot [\Theta^{(l)} h_i^{(l)} \| \Theta^{(l)} h_j^{(l)}]))}{\sum_{k \in \mathcal{N}(i)} \exp(\mathrm{LeakyReLU}(W_a^T \cdot [\Theta^{(l)} h_i^{(l)} \| \Theta^{(l)} h_k^{(l)}]))} \end{aligned} \tag{3}$$

where Θ is the learnable weight matrix, W_a is the learnable attention weight vector, LeakyReLU is the nonlinear activation function, and a_{ij} is a pairwise unnormalized attention score.

In this study, the GAT method in eq 2 was expanded for molecular graph regression tasks, specifically to predict the redox potential. This was accomplished by concatenating node (atom) representations in Table 1 with n-dimensional edge attributes e_{ij} in Table 2. Edge attributes are structurally similar to node characteristics except that they specify the nature of the edge between two nodes. As a result, the edge feature representation e_{ij} is concatenated to each node representation h_j and multiplied by the edge update matrix $\Theta^{(l)} h_j^{(l)} \rightarrow (\Theta^{(l)} h_j^{(l)}) | e_{ij}$ to produce the extended GAT model for molecular graphs (MolGAT) as

$$\begin{aligned} \alpha_{ij}^{(l)} &= \mathrm{softmax}(a_{ij}^{(l)}) \\ &= \frac{\exp(a_{ij}^{(l)})}{\sum_{k \in \mathcal{N}(i)} \exp(a_{ik}^{(l)})}, e_{ij}^{(l)} \\ &= \frac{\exp(\mathrm{LeakyReLU}(W_a^T \cdot [\Theta^{(l)} h_i^{(l)} \| (\Theta^{(l)} h_j^{(l)}) \| e_{ij}^{l}]))}{\sum_{k \in \mathcal{N}(i)} \exp(\mathrm{LeakyReLU}(W_a^T \cdot [\Theta^{(l)} h_i^{(l)} \| (\Theta^{(l)} h_k^{(l)}) \| e_{ij}^{l}]))} \end{aligned} \tag{4}$$

https://doi.org/10.1021/acsomega.3c01295
ACS Omega 2023, 8, 24268−24278

Table 1. Atom Features in a Molecular Graph Utilized to Train the MolGAT Model

Atom Features	Descriptions	Size
Atom type	Type of atoms (e.g., H, C, O, N) as one-hot vector	69
Atomic number	Atomic number of elements (e.g., H, C, O, N, Br)	1
Formal charge	Electrical charges encoded as [−3, −2, −1, 0, 1, 2, 3]	1
Radical electrons	Free-radical electrons encoded as [0, 1, 2, 3, 4]	1
Valence (neighbors)	Number of maximum valence electrons [0, 1, 2, ... 0.7]	8
Chirality	Chirality (nonsuperposable) as one-hot vector [1, 0]	1
Chirality type	Charality type (*R*, *S*) as one-hot vector 1, 0	1
Number of H's	Connected hydrogen encoded as 0, 1, 2, 3, 4	5
Hybridization	Hybridization as s, sp, sp^2, sp^3, sp^3d, sp^3d^2	7
Aromatic	An aromatic system as one-hot vector 1, 0	2
Atomic mass	Scaled atomic mass encoded as (mass − 1.008)/237.021	1
Vdm radius	Scaled Van der Waals radius for atomic volume (RVDW − 1.2)/1.35	1
Covalent radius	Scaled covalent radius encoded as $(R_{covalent} - 0.23)/1.71$	1
		99

Table 2. Bond Features in Molecular Graph Utilized to Train MolGAT Model

Bond Feature	Description	Size
Bond type	Single, double, triple, or aromatic	4
Conjugated bond	Conjugated bonds as one-hot vector	1
Ring	Bond in a ring as one-hot vector	1
Stereo	StereoNone, StereoAny, StereoZ, StereoE as one-hot vector	6
		12

The molecular data normally come in the sequential form of labeled SMILES strings. With the PyTorch Geometric[63] framework and RDKit[64] library these SMILES strings were optimally converted into a structured molecular graph object that can be used as an input for a GNN as indicated in the graph-encoding section of Figure 2. This molecular graph encoding process might be divided into three major parts, which are as follows: The first step was to develop a function that converts an RDKit atom object to the proper atom feature vector. Additionally, a technique changes an RDKit Bond object into the proper bond feature vector. Lastly, a function takes a list of SMILES strings and associated labels as inputs and creates outputs of a list of labeled Pytorch Geometric Graph objects by combining the results of the atom-feature vector and bond-feature vector from the previous steps. Graph-level classification or regression tasks, in contrast to node classification tasks, must focus on the global information present in graphs, requiring the utilization of a graph pooling technique to extract the global information. Therefore, after applying a series of three convolution layers in second section of Figure 2, the node latent representation aggregation was applied by stacking global max pooling (GMP) and global average pooling (GAP), both of which are permutation-invariant functions. After these global representation aggregation, the latent representation is supplied to multilayer perception (MLP) to predict the final target variable.

Overfitting is a common issue that arises when a model performs well on the training data but performs poorly on new data, and it is a significant challenge in machine learning. A significant portion of the literature on machine learning focuses on the development of methods to prevent overfitting. Overfitting is also a frequent occurrence during neural network training.[65] There are several ways to avoid overfitting, but for this study, batch normalization and dropout were used. Various normalization techniques are used by many different kinds of neural networks to help accelerate and/or stabilize the training process.[66] The model is atom-centric; therefore, each atom has a set of neighbor attributes that combine the features of nearby atoms and connecting bonds. As a result, linear transformation and nonlinear activation were carried out to equalize the vector length, which is notable since the vectors of the atomic features and the nearby atomic features do not have the same length. However, stacking extra layers to a GCN causes the classic vanishing gradient problem, in which back-propagating across these networks causes oversmoothing, resulting in the features of graph vertices converging to the same values.[67] Due to these limitations, the majority of cutting-edge GCNs are no deeper than four layers.[68] Hence, we employed three layers of graph convolution with a dropout of 0.2, one batch normalization at the end of the convolutions, three layers of MLP and RELU activation function in each convolution, and MLP linear years. In addition, for both training and testing, we employed a squared-error loss of the following form

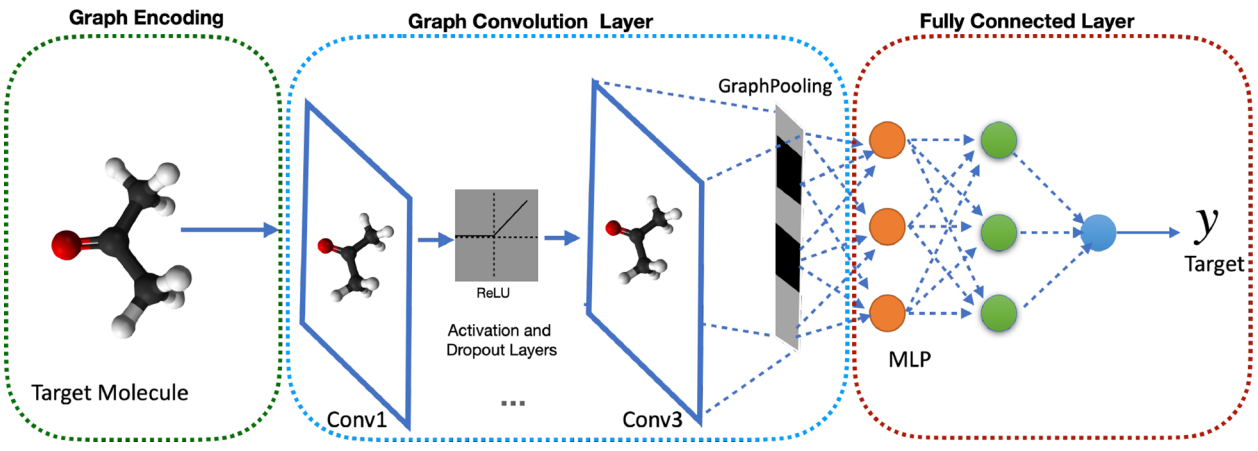


Figure 2. Schematic representation of MolGAT workflow.

https://doi.org/10.1021/acsomega.3c01295
ACS Omega 2023, 8, 24268−24278

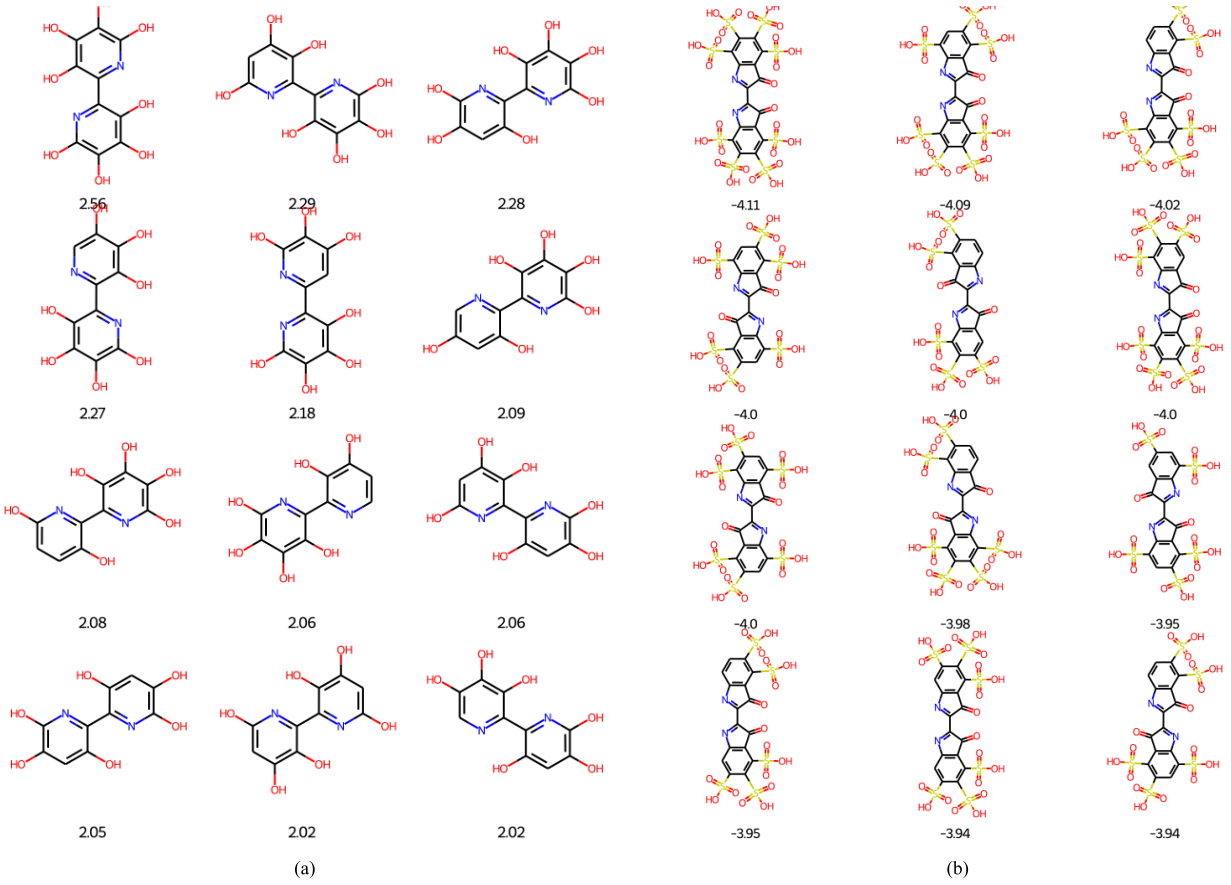


Figure 3. Some of the anolyte and the catholyte molecules with their reaction energy from RedDB: (a) catholyte materials and (b) anolyte materials.

$$L = \sum_{G_i \in T} \left\| \mathrm{MLP}(Z_{G_i}) - y_{G_i} \right\|_2^2 \tag{5}$$

where MLP is a densely connected neural network with an output and $y_{Gi} \in R$ is the target attribute value for the training graph G_i and labeled training graphs $T = \{G_1, ..., G_n\}$.

2.2. Data Preparation. To train the models RedDB[69] was used, which is a computational database that covers a chemical space of two classes of organic molecules such as quinones and aza-aromatics that are extremely promising for aqueous redox flow batteries. This data set was generated using simulation tools that use cheminformatics, machine learning, molecular mechanics, and quantum chemistry methods. It contains structural information as well as several physicochemical properties of molecules that may be used as electroactive materials in aqueous redox flow batteries.[69] The validity of the SMILES was verified using the rdkit library, and a corresponding molecular graph was generated. In this data set, redox potentials for anolytes and catholytes typically range from −4.11 to 2.56 were selected as illustrated in Figure 3 from a total of 15,000 molecules.

After collecting the data set, each compound was represented as a graph whose nodes corresponded to atoms and edges corresponded to bonds. A bond between two atoms indicates that the atoms are linked via covalent bonding. Atom types were encoded as binary bits indicating whether the atom type is carbon, hydrogen, oxygen, sulfur, phosphorus, or nitrogen. With atoms as nodes and chemical bonds as edges, each graph represents a molecule and all its structural information, atomic properties, and edge features are encoded from their smile strings using RDKit.[64] The n-dimensional atomic features used in the molecular graph structure are summarized in Table 1:

In order to include as much information as possible within the molecular graph, a wide range of atomic characteristics are included, such as atom type, formal charge, number of valence electrons, radical electrons, hybridization type, whether the atom is in a ring, whether it is aromatic, atomic mass, Van der Waals radius, and covalent radius. Due to the fact that the final three qualities are numerical in nature, they are automatically scaled using estimates based on empirical data to a respectable range. The stereochemical characteristic charity is utilized, and hydrogen is defined to be explicitly addressed in the molecular graph.

The n-dimensional bond attributes indicated in Table 2 include bond types, whether the bond is conjugated, and whether the bond is in a ring including $E-Z$ stereochemical features. The bond types are encoded as integers corresponding to the following numbers: single, double, triple, or aromatic.

2.3. Training the Models. MolGAT model was implemented using PyTorch and PyTorch Geometric (PyG)[63] Libraries. Three MolGATConv layers were used in the training, each with four attention heads, 512 hidden channels with a ReLU activation function, and a 0.2 dropout rate. In the first convolution, 99 node features were used as input and 12 edge attributes were concatenated in each convolution layer operation, with batch normalization of 512

https://doi.org/10.1021/acsomega.3c01295
ACS Omega 2023, 8, 24268−24278

hidden dimensions applied after the last convolution layer. After the last convolution layer, batch normalization of 512 hidden dimensions was applied, followed by a permutation invariant function, global max pooling (GMP), and global average pooling (GAP) to generate a graph-level embedding. These pooling operations aggregated information from all nodes in the graph to generate a single vector, effectively summarizing graph-level information. The final step was to use three fully connected layers with 1024 hidden dimensions, and ReLU activation on the first two layers was used at the end of these global aggregations. The model contains a total of 5,219,985 parameters and produces a scalar output from the final fc_out layer used for predicting the redox potential and is relatively large, as shown in Table 3.

Table 3. Trainable Parameters in MolGAT Model for Redox Potential Prediction

Layer	Shape	No. of Trainable Parms
MolGATConv1 (heads = 4)	(99, 512, 12, 4)	475696
MolGATConv2 (heads = 4)	(512, 512, 12, 4)	1321520
MolGATConv3 (heads = 4)	(512, 512, 12, 4)	1321520
BatchNorm	(512, 512)	1024
Linear1	(1024, 1024)	1049600
Linear2	(1024, 1024)	1049600
Linear3(fc_out)	(1024, 1)	1025
Total Trainable parms	5, 219, 985	5219985

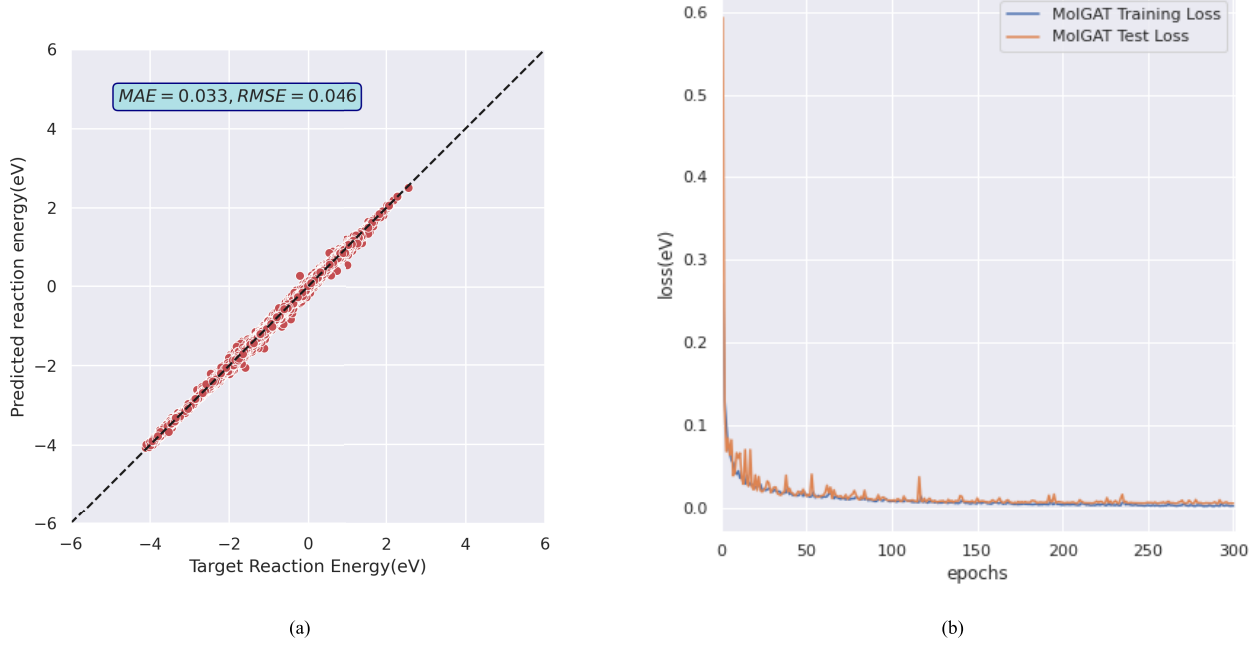
From the set of RedDB data sets, 90% was used for training and 10% for testing to training the models. The Noam Learning rate schedule was applied during the training. This corresponds to increasing the learning rate linearly for the first warmup step training steps and decreasing it thereafter proportionally to the inverse square root of the step number, scaled by the inverse square root of the dimension of the model. This learning rate increases linearly from the initial learning rate to the maximum learning rate throughout the first warmup steps. Then the learning rate decreases exponentially from the maximum learning rate to the final learning rate throughout the remaining based on the scheduling technique of Vaswani et al.[70] The training was performed with GPUs on BNL HPC for 300 epochs with a batch size of 192, an initial learning rate of $1e^{-4}$, the maximum learning rate of $1e^{-3}$, and a final learning rate of $1e^{-4}$ with warmup epochs of 2.0 based on a Noam learning rate scheduler with Adam SGD optimizer. The source code of MolGAT, the training scripts, and data sets are found at https://github.com/mesfind/molgnn.

3. RESULTS AND DISCUSSION

The MolGAT model is a powerful tool in the field of molecular property prediction. Its architecture is designed to take advantage of the graphlike nature of molecules by using graph attention networks to process the molecular graph. With a total of 6,127,249 trainable parameters, the model is capable of capturing complex relationships between atoms and their surroundings as well as between different parts of the molecule.

The three MolGATConv layers in the model are responsible for performing graph convolution operations with each layer gradually increasing the number of output channels while maintaining the same number of heads. The heads parameter refers to the number of parallel attention mechanisms used to compute the node representations. The output shape of each convolutional layer reflects the fact that the model is capable of processing both local and global information with the output channels representing different levels of abstraction.

The three linear layers in the model are fully connected layers responsible for performing the final processing of the node and edge features. The first two linear layers have 1024 output features each, allowing for additional nonlinear transformations of the features learned by the convolutional layers. The final linear layer is the output layer of the model



(a) (b)

Figure 4. MolGAT training performance: (a) parity plot of predicted against target values and (b) MolGAT model's test and training losses.

https://doi.org/10.1021/acsomega.3c01295
ACS Omega 2023, 8, 24268−24278

with a single output feature representing the predicted molecular property.

3.1. Predicting Target. The Adam optimizer, an extension of stochastic gradient descent with minibatches normalization, enables the MolGAT model to manage a noisy optimization of the high-dimensional nature of the network's weights. The training's mean average error (MAE) and root-mean-square error (RMSE) are 0.033 *V* and 0.046 *V* respectively as shown in Figure 4a and Figure S7. To enhance the performance of the MolGAT model, the redox potential values were standardized by computing the mean and standard deviation of the data and then using the formula (data − mean)/std to transform the target variable to have a mean of 0 and standard deviation of 1. Following training, the predictions were transformed back using the inverse formula (data × std) + mean. Standardizing the redox potential also facilitated a comparison of the performance of different GNN models and interpretation of their results.

The parity plot in Figure 4 revealed a close agreement between predicted and targe values, indicating the reliability of the MolGAT model's predictions. Furthermore, the model's test and training losses were also calculated, which showed that the model was not overfitting the data and was effectively learning the patterns in the data.

The performance of various models on the RedDB data set was compared after standardizing the data and training for 300 epochs using similar training procedures The performance comparison (Table 4) shows the MAE and RMSE for different models on the RedDB data set.

Table 4. Model Performance Comparison Using the RedDB Dataset

Model	MAE	RMSE
MPNN	0.083	0.117
GCN	0.075	0.101
GAT	0.061	0.086
AttentiveFP	0.060	0.086
MolGAT (this study)	0.033	0.046

The results shows the MAE and RMSE for different models, namely MPNN, GCN, GAT, AttentiveFP, and MolGAT (this study, Table 4, Figure S8). It is evident that the MolGAT model has lower MAE and RMSE values, indicating its superiority over all of the other models. Even though MPNN and GCN had errors higher than those of GAT and AttentiveFP, they still outperformed the MolGAT model. These results imply that MolGAT is more effective in predicting the redox potential of molecules in the RedDB data set. Thus, the MolGAT model is a promising tool for predicting other molecular properties such as solubility, making it a useful tool in redox flow battery, drug discovery, and other applications.

3.2. Virtual Screening (VS) Promising Redox-Active Molecules. High-throughput screening (HTS) is a process used to rapidly test a large number of compounds or materials for their desired properties. Graph neural network (GNN) models have recently emerged as a promising tool for HTS due to their ability to learn from graph-structured data, such as molecular structures. GNN models can be trained on a set of labeled compounds or materials with known properties, allowing them to learn the relationships between the molecular structure and property. Once trained, the model can be used to predict the properties of a large set of unlabeled compounds or materials, enabling researchers to quickly screen for the desired properties. One advantage of using GNN models for HTS is that they can capture complex relationships between atoms and molecules, which can be difficult to model with traditional machine learning methods.

The use of GNN also presents an exciting opportunity for accelerating the discovery of new materials for redox flow applications through virtual screening to identify the desired properties. In addition, the GNN enables the exploration of previously undiscovered chemical space, increasing the number of libraries that can be screened by those in the field. Specifically, the molecular graph attention network (MolGAT) model, which is a type of GNN, predicts the redox potential of organic compounds by leveraging fundamental chemical properties and using molecular graphs for representation learning with both structural and chemical features. The MolGAT model was used to screen 581,014 molecules from various databases and identified promising catholytes and anolytes based on their redox potential values. The MolGAT model was used to calculate the redox potential of 581,014 molecules from various databases such as ZINC,[24] CHEMBL,[23] DELANEY,[36] OMDB,[71] and QM9[72] and identified promising catholytes and anolytes based on their redox potential values. A data set containing 23,467 molecules with redox potentials greater than 1 V and less than − 1 V was created, with 20,716 molecules identified as promising catholytes and 2,751 molecules identified as promising anolytes. From this data set, 12 molecules with promising redox potential values were chosen for illustration from both catholyte in Figure 5a and anolyte in Figure 5b. This approach provides a new way to identify potential candidates for electrochemical storage systems, and their redox potentials can be improved by adding substituents to the base molecules Figure S6 using DFT methods in future research. The use of the MolGAT model for screening redox potential values offers a powerful tool for accelerating the discovery of new materials with desirable properties.

MolGAT is a graph-based deep learning model that utilizes attention-based message passing operations to extract important features and relationships between atoms from molecular structure input. The model is trained to predict the redox potential of organic compounds using molecular structures, atomic attributes, and bond attributes. This makes it an ideal tool for virtual screening, which is a process that involves screening large databases of compounds or materials to identify those with the desired properties.

In this case, virtual screening was performed using the MolGAT model to screen several redox-active molecules from various databases, such as ZINC, DELANEY, QM9, OMDB, and CHEMBL. The trained model was able to predict the redox potential of these molecules accurately, as demonstrated in Table 4. The performance of the MolGAT model was compared to other GNN models such as GCN, GAT, and AttentiveFP, and the results showed that the MolGAT model had superior generalization performance. This means that the MolGAT model was able to accurately predict the redox potential of molecules that it had not been trained on. This is a crucial characteristic for high throughput virtual screening, as it allows researchers to quickly and accurately screen large databases of compounds or materials to identify those with desired properties.

https://doi.org/10.1021/acsomega.3c01295
ACS Omega 2023, 8, 24268−24278

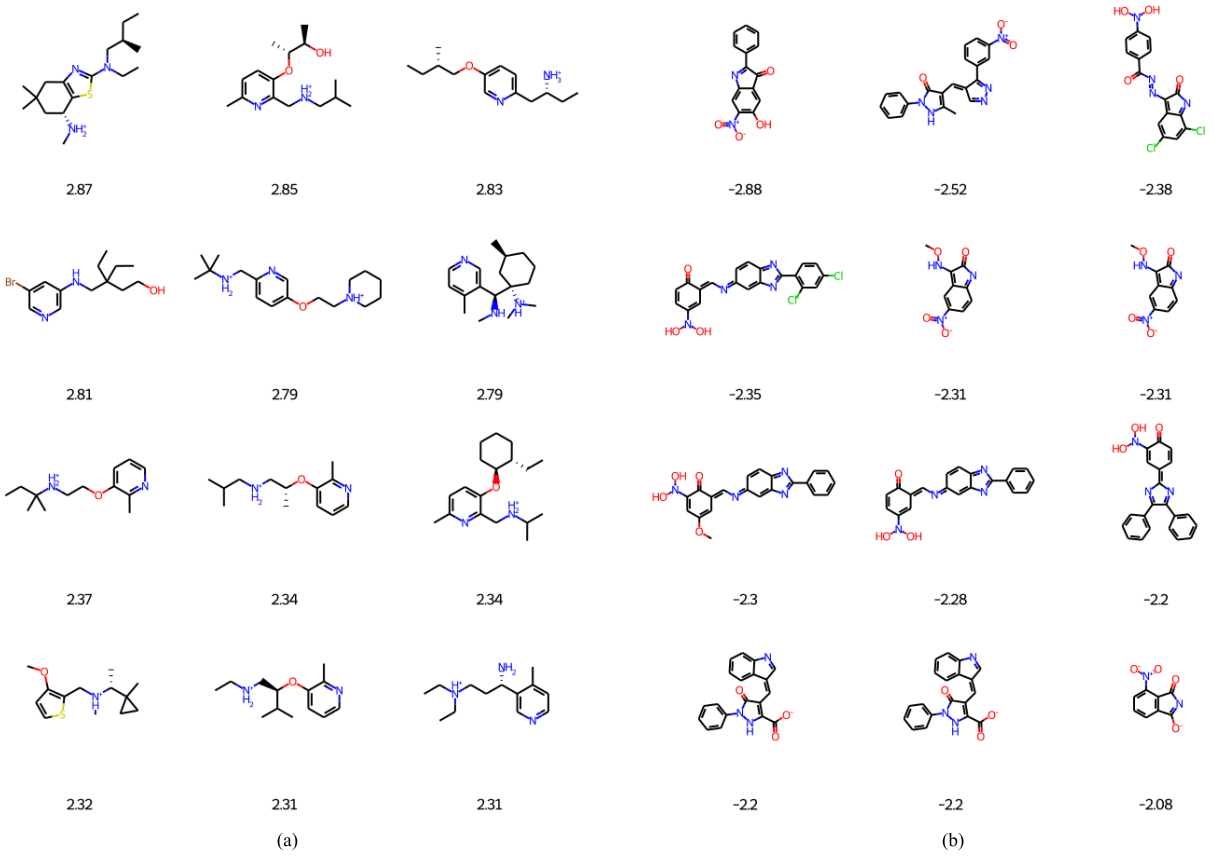


Figure 5. Some selected promising redox-active molecules with their corresponding predicted redox potential values screened with the MolGAT model: (a) catholyte materials and (b) anolyte materials.

4. CONCLUSION

The MolGAT model is a powerful tool for predicting molecular properties by using a graph-based deep-learning approach. By leveraging attention-based message passing operations, the model can effectively extract crucial features and relationships between atoms from the input molecular structure. Importantly, the model's ability to learn the chemical structure and properties, coupled with its consideration of both structural and chemical features, enables it to accurately predict molecular properties. This eliminates the need for complex feature engineering, which is often time-consuming and labor-intensive in traditional machine learning approaches. This information is then processed through fully connected layers to predict the redox potential of the organic compounds.

The trained model was used as a virtual screening tool to identify promising redox-active organic compounds from various databases to accelerate the discovery of new materials for redox flow battery applications. These screened compounds may be further optimized by adding functional groups to the base molecules using forward modeling in DFT or experimental approaches in future research. Moreover, the proposed method can be extended to predict additional chemical attributes such as solubility and stability, allowing for a more targeted screening of materials. The importance of considering edge attributes in GNN models when predicting molecular properties is also emphasized. In general, the MolGAT model demonstrates the potential of graph neural networks in condensed matter physics and materials science. Its ability to predict molecular properties with appropriate fundamental atomic and edge features inputs, combined with its fast and reliable prediction capabilities, opens up a new field of virtual screening to identify materials with desirable properties. Additionally, the model's applicability extends beyond the realm of battery research to the general screening of promising materials from large databases.

■ ASSOCIATED CONTENT

Data Availability Statement

The RedDB data set used for this work was collated from the openly available harvard dataverse studied in ref 69. The MolGAT model screened data set redox-active organic compounds are available for further investigation using electronic structure methods like DFT or experimental techniques.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c01295.

Explanation of the data preprocessing steps undertaken to train the MolGAT model; extensive analysis of the RedDB data set, including joint plots of reaction energy and number of atoms, box plots of the number of atoms, the number of heavy atoms in anolyte and catholyte molecules, and the types of functional groups present in both catholyte and anolyte; a list of the sample SMILES strings with their corresponding redox potentials in the RedDB data set used for training, along with instructions on how to load this data set in PyG library graph representation format; normalization of the target variable (redox potential) and sample atomic features

https://doi.org/10.1021/acsomega.3c01295
ACS Omega 2023, 8, 24268−24278

and edge attributes used to train the MolGAT model; outline of the error loss plots of the different graph neural network models used in the study, including MolGAT, GCN, GAT, AttentiveFP, and MPNN, along with their corresponding parity plots with MAE and RMSE loss values for comparison; a list of the top-performing molecules selected from the screened data set, along with their corresponding number of heavy atoms, types of functional groups, and other descriptors (PDF)

AUTHOR INFORMATION

Corresponding Author

Yedilfana Setarge Mekonnen − *Center for Environmental Science, College of Natural and Computational Sciences, Addis Ababa University, Addis Ababa 1176, Ethiopia;* orcid.org/0000-0001-9331-7370; Email: yedilfana.setarge@aau.edu.et, mesfin.diro@aau.edu.et

Authors

Mesfin Diro Chaka − *Department of Physics, College of Natural and Computational Sciences, Addis Ababa University, Addis Ababa 1176, Ethiopia; Computational Data Science, College of Natural and Computational Sciences, Addis Ababa University, Addis Ababa 1176, Ethiopia;* orcid.org/0000-0003-4248-9644

Chernet Amente Geffe − *Department of Physics, College of Natural and Computational Sciences, Addis Ababa University, Addis Ababa 1176, Ethiopia*

Alex Rodriguez − *The Abdus Salam International Centre for Theoretical Physics(ICTP), Condensed Matter and Statistical Physics Section, 34100 Trieste, Italy;* orcid.org/0000-0002-0213-6695

Nicola Seriani − *The Abdus Salam International Centre for Theoretical Physics(ICTP), Condensed Matter and Statistical Physics Section, 34100 Trieste, Italy;* orcid.org/0000-0001-5680-9747

Qin Wu − *Brookhaven National Laboratory, Center for Functional Nanomaterials, Upton, New York 11973, United States;* orcid.org/0000-0001-6350-6672

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c01295

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by a thematic research project (Grant No. TR/036/2020) sponsored by Addis Ababa University. Computer resources for the study were kindly provided by the Center for Functional Nanomaterials (CFN), which is a U.S. Department of Energy Office of Science User Facility, at Brookhaven National Laboratory under Contract No. DE-SC0012704 and Project No. 308379 and the Ethiopian Education and Research Network (EthERNet) at the Ethiopian Ministry of Education. Moreover, Y.S.M. would like to acknowledge support from the ICTP through the Associates Programme (2020−2025).

REFERENCES

(1) Dunn, B.; Kamath, H.; Tarascon, J.-M *Science* **2011**, *334* (6058), 928−935.
(2) Nitta, N.; Wu, F.; Lee, J. T.; Yushin, G *Materials Today* **2015**, *18* (5), 252−264.
(3) Tarascon, J.-M.; Armand, M *Nature* **2001**, *414* (6861), 359−367.
(4) Mekonnen, Y. S.; Christensen, R.; Garcia-Lastra, J. M.; Vegge, T *The Journal of Physical Chemistry Letters* **2018**, *9* (15), 4413−4419.
(5) Mekonnen, Y. S.; Knudsen, K. B.; Mýrdal, J. S. G.; Younesi, R.; Højberg, J.; Hjelm, J.; Norby, P.; Vegge, T *The Journal of Chemical Physics* **2014**, *140* (12), 121101.
(6) Benti, N. E.; Gurmesa, G. S.; Geffe, C. A.; Mohammed, A. M.; Tiruye, G. A.; Mekonnen, Y. S *J. Mater. Chem. A* **2022**, *10* (15), 8501−8514.
(7) Lee, B.-J.; Zhao, C.; Yu, J.-H.; Kang, T.-H.; Park, H.-Y.; Kang, J.; Jung, Y.; Liu, X.; Li, T.; Xu, W.; Zuo, X.-B.; Xu, G.-L.; Amine, K.; Yu, J.-S. *Nature Communications* **2022**, *13* (1), 4629 DOI: 10.1038/s41467-022-31943-8.
(8) Gurmesa, G. S.; Benti, N. E.; Chaka, M. D.; Tiruye, G. A.; Zhang, Q.; Mekonnen, Y. S.; Geffe, C. A *RSC Advances* **2021**, *11* (16), 9721−9730.
(9) Sakata Gurmesa, G.; Teshome, T.; Ermias Benti, N.; Ayalneh Tiruye, G.; Datta, A.; Setarge Mekonnen, Y.; Amente Geffe, C. *ChemistryOpen* **2022**, *11* (6), e202100289 DOI: 10.1002/open.202100289.
(10) Zhong, F.; Yang, M.; Ding, M.; Jia, C *Front. Chem.* **2020**, *8*, 451.
(11) Ye, W. C *Nat Commun* **2022**, *13* (1), 3184.
(12) Sánchez-Díez, E.; Ventosa, E.; Guarnieri, M.; Trovó, A.; Flox, C.; Marcilla, R.; Soavi, F.; Mazur, P.; Aranzabe, E.; Ferret, R *Journal of Power Sources* **2021**, *481*, 228804.
(13) Huskinson, B.; Marshak, M. P.; Suh, C.; Er, S.; Gerhardt, M. R.; Galvin, C. J.; Chen, X.; Aspuru-Guzik, A.; Gordon, R. G.; Aziz, M. J *Nature* **2014**, *505* (7482), 195−198.
(14) Er, S.; Suh, C.; Marshak, M. P.; Aspuru-Guzik, A *Chem. Sci.* **2015**, *6* (2), 885−893.
(15) Chai, L. J. *Electroactive materials for next-generation redox flow batteries: From inorganic to organic*; American Chemical Society, 2020; Vol. *1364*, pp 1−47.
(16) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H; et al. *J. Med. Chem.* **2020**, *63* (16), 8749−8760.
(17) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F *Chemical Science* **2019**, *10* (2), 370−377.
(18) Goodall, R. E. A.; Lee, A. A. *Nat Commun* **2020**, *11* (1), 6280.
(19) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C *JOM* **2013**, *65* (11), 1501−1509.
(20) Cheng, L.; Assary, R. S.; Qu, X.; Jain, A.; Ong, S. P.; Rajput, N. N.; Persson, K.; Curtiss, L. A *J. Phys. Chem. Lett.* **2015**, *6* (2), 283−291.
(21) Wang, J.; Wang, Y.; Chen, Y *Materials* **2022**, *15* (5), 1811.
(22) Pelzer, K. M.; Cheng, L.; Curtiss, L. A *The Journal of Physical Chemistry C* **2017**, *121* (1), 237−245.
(23) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibri'an-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R *Nucleic Acids Res.* **2017**, *45* (D1), D945−D954.
(24) Irwin, J. J.; Shoichet, B. K *J. Chem. Inf. Model.* **2005**, *45* (1), 177−182.
(25) Roy, D.; Mandal, S. C.; Pathak, B *ACS Appl. Mater. Interfaces* **2021**, *13* (47), 56151−56163.
(26) Saeki, A.; Kranthiraja, K. *Jpn. J. Appl. Phys.* **2020**, *59*, SD0801.
(27) Song, S.; Wang, Y.; Chen, F.; Yan, M.; Zhang, Q *Engineering* **2022**, *10*, 99−109.
(28) Zafari, M.; Kumar, D.; Umer, M.; Kim, K. S *J. Mater. Chem. A* **2020**, *8* (10), 5209−5216.

https://doi.org/10.1021/acsomega.3c01295
ACS Omega 2023, 8, 24268−24278

(29) Allam, O.; Kuramshin, R.; Stoichev, Z.; Cho, B. W.; Lee, S. W.; Jang, S. S *Materials Today Energy* **2020**, *17*, 100482.
(30) Mueller, T.; Kusne, A. G.; Ramprasad, R. *Reviews in Computational Chemistry*; John Wiley & Sons, Inc., 2016; pp 186−273.
(31) Yamada, H.; Liu, C.; Wu, S.; Koyama, Y.; Ju, S.; Shiomi, J.; Morikawa, J.; Yoshida, R *ACS Cent. Sci.* **2019**, *5* (10), 1717−1730.
(32) Segler, M. H. S.; Preuss, M.; Waller, M. P *Nature* **2018**, *555* (7698), 604−610.
(33) Agrawal, A.; Choudhary, A *APL Mater.* **2016**, *4* (5), 053208.
(34) Ryan, K.; Lengyel, J.; Shatruk, M *J. Am. Chem. Soc.* **2018**, *140* (32), 10158−10168.
(35) Deffrennes, G.; Terayama, K.; Abe, T.; Tamura, R *Materials & Design* **2022**, *215*, 110497.
(36) Delaney, J. S *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 1000−1005.
(37) Wei, J.; Chu, X.; Sun, X.-Y.; Xu, K.; Deng, H.-X.; Chen, J.; Wei, Z.; Lei, M *InfoMat* **2019**, *1* (3), 338−358.
(38) Chen, C.-T.; Gu, G. X *Adv. Sci.* **2020**, *7* (5), 1902607.
(39) Kim, K.; Kang, S.; Yoo, J.; Kwon, Y.; Nam, Y.; Lee, D.; Kim, I.; Choi, Y.-S.; Jung, Y.; Kim, S.; Son, W.-J.; Son, J.; Lee, H. S.; Kim, S.; Shin, J.; Hwang, S *npj Comput Mater* **2018**, *4* (1), 67.
(40) Noh, J.; Kim, J.; Stein, H. S.; Sanchez-Lengeling, B.; Gregoire, J. M.; Aspuru-Guzik, A.; Jung, Y *Matter* **2019**, *1* (5), 1370−1384.
(41) Wang, Q.; Zhang, L *Nat Commun* **2021**, *12* (1), 5359.
(42) Yao, Z.; Sanchez-Lengeling, B.; Bobbitt, N. S.; Bucior, B. J.; Kumar, S. G. H.; Collins, S. P.; Burns, T.; Woo, T. K.; Farha, O.; Snurr, R. Q.; Aspuru-Guzik, A. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nature Journal* **2021**, *3*, 76.
(43) Huo, H.; Rupp, M. *arXiv: Chemical Physics* **2017**, 1704.06439 DOI: 10.48550/arXiv.1704.06439.
(44) Rogers, D.; Hahn, M *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.
(45) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional networks on graphs for learning molecular fingerprints. *ArXiv* **2015**, DOI: 10.48550/arXiv.1509.09292.
(46) Morgan, H. L *J. Chem. Doc.* **1965**, *5* (2), 107−113.
(47) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V *J. Chem. Inf. Model.* **2015**, *55* (2), 263−274.
(48) Hamilton, W. L.In *Graph representation learning*; Springer International Publishing: Cham, 2022; pp 51−70.
(49) Bresson, X.; Laurent, T. Residual gated graph ConvNets. *ArXiv* **2018**, DOI: 10.48550/arXiv.1711.07553.
(50) Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *ArXiv* **2017**, DOI: 10.48550/arXiv.1609.02907.
(51) Kojima, R.; Ishida, S.; Ohta, M.; Iwata, H.; Honma, T.; Okuno, Y *J Cheminform* **2020**, *12* (1), 32.
(52) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P *J Comput Aided Mol Des* **2016**, *30* (8), 595−608.
(53) Na, G. S.; Kim, H. W *Neural Networks* **2021**, *133*, 1−10.
(54) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M *J. Med. Chem.* **2020**, *63* (16), 8749−8760.
(55) Jiang, D.; Wu, Z.; Hsieh, C.-Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T *J Cheminform* **2021**, *13* (1), 12.
(56) Hamilton, W. L *Synthesis Lectures on Artificial Intelligence and Machine Learning* **2020**, *14* (3), 1−159.
(57) Xie, T.; Grossman, J. C *Phys. Rev. Lett.* **2018**, *120* (14), 145301.
(58) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. *CoRR* **2017**, DOI: 10.48550/arXiv.1704.01212.
(59) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P *Chem. Mater.* **2019**, *31* (9), 3564−3572.
(60) Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; van den Berg, R.; Titov, I.; Welling, M. Modeling Relational Data with Graph Convolutional Networks. *ArXiv* **2017**, DOI: 10.48550/arXiv.1703.06103.
(61) Gong, L.; Cheng, Q. *CoRR* **2018**, DOI: 10.48550/arXiv.1809.02709.
(62) Veličkovi'c, P.; Cucurull, G.; Casanova, A.; Romero, A.; Li'o, P.; Bengio, Y. *ArXiv* **2018**, DOI: 10.48550/arXiv.1710.10903.
(63) Fey, M.; Lenssen, J. E. *ArXiv* **2019**, DOI: 10.48550/arXiv.1903.02428.
(64) Landrum, G.; Tosco, P.; Kelley, B.; Ric; sriniker; gedeck; Vianello, R.; Schneider, N.; Kawashima, E.; Dalke, A.; N, D.; Cosgrove, D.; Cole, B.; Swain, M.; Turk, S.; Savelyev, A.; Jones, G.; Vaucher, A.; Wójcikowski, M.; Take, I.; Probst, D.; Ujihara, K.; Scalfani, V. F.; Godin, G.; Pahl, A.; Berenger, F.; Varjo, J. L.; strets123, J. P.; Doliath, G. *Rdkit/rdkit: 2022_03_1 (Q1 2022) release*; Zenodo, 2022.
(65) Pyzer-Knapp, E. O.; Li, K.; Aspuru-Guzik, A. *Adv. Funct. Mater.* **2015**, *25* (41), 6495−6502.
(66) Cai, T.; Luo, S.; Xu, K.; He, D.; Liu, T.-Y.; Wang, L. GraphNorm: A Principled Approach to Accelerating Graph Neural Network Training. *ArXiv* **2021**, DOI: 10.48550/arXiv.2009.03294.
(67) Li, Q.; Han, Z.; Wu, X.-M. *CoRR* **2018**, DOI: 10.48550/arXiv.1801.07606.
(68) Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph Neural Networks: A Review of Methods and Applications. *ArXiv* **2021**, DOI: 10.48550/arXiv.1812.08434.
(69) Sorkun, E.; Zhang, Q.; Khetan, A.; Sorkun, M. C.; Er, S. *ChemRxiv* **2021**, DOI: 10.26434/chemrxiv.14398067.v1.
(70) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *ArXiv* **2017**, DOI: 10.48550/arXiv.1706.03762.
(71) Borysov, S. S.; Geilhufe, R. M.; Balatsky, A. V. *PLOS ONE* **2017**, *12* (2), No. e0171501.
(72) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Lilienfeld, O. A. von. *Scientific Data* **2014**, *1*, 1 DOI: 10.1038/sdata.2014.22.

Recommended by ACS

DFRscore: Deep Learning-Based Scoring of Synthetic Complexity with Drug-Focused Retrosynthetic Analysis for High-Throughput Virtual Screening

Hyeongwoo Kim, Woo Youn Kim, *et al.*

AUGUST 31, 2023
JOURNAL OF CHEMICAL INFORMATION AND MODELING READ

pBRICS: A Novel Fragmentation Method for Explainable Property Prediction of Drug-Like Small Molecules

Sarveswara Rao Vangala, Arijit Roy, *et al.*

AUGUST 16, 2023
JOURNAL OF CHEMICAL INFORMATION AND MODELING READ

FFLOM: A Flow-Based Autoregressive Model for Fragment-to-Lead Optimization

Jieyu Jin, Yu Kang, *et al.*

JULY 20, 2023
JOURNAL OF MEDICINAL CHEMISTRY READ

Equivariant Graph Neural Networks for Toxicity Prediction

Julian Cremer, Gianni De Fabritiis, *et al.*

SEPTEMBER 10, 2023
CHEMICAL RESEARCH IN TOXICOLOGY READ

Get More Suggestions >

https://doi.org/10.1021/acsomega.3c01295
ACS Omega 2023, 8, 24268−24278