**Electrolytes for high-voltage Lithium-ion battery: A new approach with machine learning**

Ming-Hsiu Wu,1 Chih-Ao Liao,1 Ngoc Thanh Thuy Tran,12 Wen-Dung Hsu12

1 Department of Materials Science and Engineering, National Cheng-Kung University

2 Hierarchical Green-Energy Materials (Hi-GEM) Research Center, National Cheng Kung University

**Abstract:** Recently, high-voltage Lithium-ion batteries (LIBs) has gained a lot of attraction because of the potential application in electric vehicles.However, conventional carbonate solvent-based electrolytes are less stable against high-voltage cathodes. In this thesis, machine learning to design new electrolytes for high-voltage Lithium-ion battery have been discussed. Different metrics for molecular selection, training and predicting machine learning models are considered to find out an appropriate method. The two important parameters for characterizing electrolytes, namely electron affinity (EA) and Ionization energy (IE), are chosen. Our results show that the SMILES (one-hot encoding) format combined with recurrent neural network (RNN) to predict EA/IE has better prediction performance compared to other neural network models. Furthermore, the generative model was applied to carry out inverse design. The machine learning techniques lead to a new way to create brand new functional molecules for electrolyte without doing complicated calculations and costly experiments. The generative model could serve as a first step design towards further investigation into other necessary properties of electrolytes for high-performance LIBs.

**Introduction**

In recent years, with a high demand on electric vehicle, developing safe and efficient high-voltage lithium-ion battery become more and more urgent[1]. However, the decomposition of traditional organic electrolyte compounds during charging and discharging significantly impedes the development. Therefore, it is desired to design new electrolyte molecules that can withstand in high voltage to replace the old ones. There are usually two directions to design a new molecule, including functional group modification and machine learning method. In this chapter we will discuss the latter method. This method provides a whole new road to approach material design.



Figure 2. 1 The chemical potential of lithium-ion battery

**Metrics for molecular selection**

According to frontier molecular orbital theory, K. Fukui et al[2] have pointed out that electrochemical stable window of the electrolytes refers to the lowest occupied molecular orbital (LUMO), the highest occupied molecular orbital (HOMO) and the interval in between LUMO and HOMO. The reactivity can be predicted by comparing the electrochemical potential of the electrode (the electron Fermi level) with the LUMO and HOMO of the electrolyte molecule. The LUMO and HOMO are usually referred to the electron orbitals of a molecule in neutral state. In redox reaction, however, ability of electron accepting, or withdrawing is the direct property to determine if the reaction would occur or not. Therefore, instead of LUMO and HOMO, electron affinity (EA) and ionization energy (IE) are considered in most researches. The energy levels of EA and IE can be defined as the redox standard, and also can take the structural changes into account. In other words, the value of IE/EA can be important properties (characteristics) for electrolyte molecular selection.

For an ideal molecule for electrolyte, its IE value should be higher than work function of cathode materials to prevent oxidation by cathode materials. For high-voltage battery the work function of cathode materials is lower than conventional ones, thus a high-IE electrolyte molecule is needed as shown in Figure 2. 1. Similarly, at the anode side in order to prevent reduction by anode materials, the EA value for the electrolyte molecule should be lower than the work function of anode materials.

To do the molecular selection for developing new electrolyte, other properties, such as structural stability during reduction or oxidation, dipole moment and polarizability, are also supposed to be taken into consideration. However, because of limited molecular properties included in current databases, such as Materials Project[3], QM9[4], ZINC[5], etc., full property selection is not possible. IE and EA, the most important characteristics, were used to do the first step selection. Additionally, viscosity of electrolyte is also an important property to consider. The viscosity of molecule is highly related to its dipole moment and polarizability, so one can take these properties into account in the next step selection in the future. Furthermore, stability of molecular structure can play dominant role in molecular selection, since it indirectly but closely affects the stability and durability of battery. When molecules are oxidized or reduced by electrode, molecular bonds possibly break and then decomposes into flammable small molecules, which lead to devastating result.

**Experiments, first-principles calculation, and machine learning**

Conventional molecules design is a time-consuming, laborious, as well as expert-based mission. The most common method is trial-and-error. Countless material development experiments are often evitable, and whole process is often high investment involved, no matter in time or human power. Historically, the time scale for development of new material, from laboratory to commercial application, is about 15 to 20 years[6]. The number of possible small molecular structures are estimated to be about on the order of 1060[7], constituting well-known chemical space. It is almost impossible to explore this space via conventional trial-and-error experimental method. First-principles calculation can provide material developers another approach which circumvents the arduousness of experimentation. By quantum mechanics, physical and chemical properties of these molecules can be exactly obtained through solving Schrodinger equation. First-principles calculation strategy significantly reduces the cost of developing new molecules, though vast chemical space still cannot be efficiently exploited, and virtually only little part of molecules in the world has been investigated until now.

The conception of machine learning has existed for a long time since Arthur Samuel first came up with the phrase “Machine Learning” in 1952. However, the lack of material data significantly impeded the development of machine learning in material science field. During the past ten years, a large amount of data including material structures and its corresponding properties have been accumulated, especially from first-principles calculation. Hence, the power of machine learning has drawn people’s attention away from conventional methods. Recent years have seen a dramatic proliferation of research concerned with unveiling the relationship between material structures and their properties from existing data. Gradually, machine learning become a powerful method for investigating materials at a large scale in the initial stage of material design.

**Introduction to first-principles calculation**

First-principles calculation is a kind of method of calculating properties directly based on the law of quantum mechanics without introducing fitting parameters or empirical modeling. Starting from the Coulomb interaction between atomic nucleus, electron-atomic nucleus and electrons, this method provides detailed insight into the origin of mechanical, electronic, optical and magnetic properties of materials and molecules. Thanks to the great advancement of the computer performance, the implementation of first-principles calculation for many body systems has become more feasible with lower cost, and thus the method has been widely used for materials research. The method not only promotes a deeper understanding of existing materials, but also can predict properties of new materials for material design.

**Introduction to Density functional theory (DFT)**

The holy grail in solid state physics is the solution of the time-independent Schrödinger equation. It can be written as

*=*

is the Hamiltonian for a system including *N* electrons and *M* nuclei. a wavefunction of coordinates of all electrons and ions.

Hamiltonian can be written as

Here, *i* and *j* denote the N electrons while A and B represent the M nuclei. The first two terms are the kinetic energy of the electron and nuclei. The other three terms represent the electrostatic potential energy of nucleus- nucleus, electron-electron, and nucleus-electron interaction.

Born-Oppenheimer approximation was introduced to reduce this equation and separate this problem into the electron-structure problem and the nuclear problem.

The approximation goes that the nuclei move much slower than the electrons because of the larger mass of nucleus. We can consider the electrons moving in the field of fixed nuclei. Therefore, the nuclear kinetic energy is zero and their potential energy is merely a constant. The electronic Hamiltonian can be reduced to

The solution of the Schrödinger equation is the electronic wave function and the electronic energy. The total energy is the sum of and the constant nuclear potential term

**The variational principle for ground state**

Given a wavefunction , the expectation value of the system energy is

We can employ the variational theorem, which states that the energy computed from a guessed wavefunction is an upper bound to the true-ground state energy.

**Hartree fock theory**

Now, we need to deal with the wave function , but no one can truly describe this function. In the theory of Hartree fock, Pauli Exclusion Principle is taken into consideration, and the electron ground state wave function is approximated as an antisymmetric product of N orthonormal spin orbitals , as known Slater Determinant.

The energy (expectation value of the Hamiltonian operator with ) can be written

describe the electron kinetic energy and electron-nucleus interaction. And are coulomb term and exchange term, respectively. The coulomb operator represents the potential an electron experience from the average charge distribution of another electron. The is the exchange contribution to the HF potential.

**The Hohenberg-Kohn (HK) theorem**

Hohenberg-Kohn theorem reformulate the many-body system in terms of the density and demonstrates that the electron density uniquely determines the Hamiltonian operator and thus all the properties of the system. It proposes two theorems. The first theorem is for any electronic system an external potential is determined uniquely by the ground state electron density . The second theorem is the ground state energy can be obtained variationally which means the density that minimize the total energy is the exact ground state density. According to the Hohenberg-Kohn theorem, the total energy can be written as

The energy functional have been separated into a universal part (independent on the system) and a system dependent part . The functional describes the functional for the kinetic energy and that for the electron-electron interaction . And the functional can be decomposed into

is the classical electron-electron coulomb potential and term is the non-classical contribution to electron-electron interaction (self-interaction correction, exchange and Coulomb correlation). But, the explicit form of the functional and becomes the major challenge of density functional theory. HK-DFT is in principle exact but impractical. Later, Kohn-Sham approach attempt to solve this problem.

**The Kohn-Shan equations**

The theory assumes that ground state density of interacting system is equal to that of some non-interacting system which is soluble but including some approximation. The term could be reformulated as

where represents the kinetic energy of non-interacting particles, and , the so-called exchange-correlation is defined as

The exchange and correlation energy contains everything that is unknown. The Kohn-Sham one-particle equations is

Where is KS orbital and is KS eigen energy.



Nowadays, density functional theory (DFT) are used to calculate ground-state electronic structure of atoms, molecules and solid state materials. This theory begins from the first and second Hohenberg-Kohn theorems. The Thomas-Fermi approximation and its limitations are reviewed. The remedy for its problems are incorporated into the Kohn-Sham equations for evaluating the ground-state energy. The exchange-correlation energy contribution to the Kohn-Sham equations can be approximated by the local density approximation or several other approximations. The spin-DFT formalism is introduced, and the gap problem, related to the evaluation of the fundamental gap (the energy difference between the ionization potential and electron affinity), is discussed. Time-dependent DFT and the Runge-Gross theorem that is the analog of the Hohenberg-Kohn theorem for time-dependent external potentials are presented.

**The conception of Machine learning**

Formally, machine learning is the science of getting computers to accomplish a task without a set of rules defined by human. In field of machine learning, the underlying algorithm is selected or designed by human. The machine learning algorithms learn the knowledge or the parameters from data. In other words, a data set is used to train a mathematical model so that when it sees similar data in the future, it knows what to do with it. Models typically take data as an input and then output a prediction of something of interest. As the large amount of data has been accumulated, machine learning gain momentum and gradually become indispensable in many different fields, such as web search, computational biology, finance, robotics, information extraction, and social network. There are mainly four types of machine learning nowadays: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.

The first one is supervised learning, the machine learns from a set of data including example input and example output. Given there are enough input-output data, the machine can learn the relationship between input and output and construct a robust model. From there, the model can be used to predict unknown outputs from known input. In reality, the supervised learning problem can be divided into classification and regression. The aim of classification problem is to determine which category the given input data belongs to. The prediction output is typically discrete type. A classical example is categorizing animal picture into a cat or a dog. As for the problem of regression, the output is continuous type. For example, the model could predict someone’s age from some features. In this work, we have tried to use molecular features to predict their IE/EA. Since the value of IE/EA is continuous type, this study belongs to supervised learning regression problem. The biggest challenge of supervised learning is that it required a large amount of labeled data, but in general labeling the data will entail a large amount of manpower.

The second one is unsupervised learning. Unsupervised learning is for situation where there is only a set of input data but no output target. Therefore, we cannot make predictions. Instead, the most common mothed in unsupervised learning is cluster analysis. Its aim is to find a hidden pattern in data by grouping similar things together.

The next one is semi-supervised learning. In the most of cases the amount of labeled data is scarce, which is not enough to train a good model, and most of data is unlabeled. The ultimate point of unsupervised learning is to make use of unlabeled data to make a better prediction, not just from labeled data.

The last one is reinforcement learning. The central concept of reinforcement learning is based on an “agent” (a computer or robot) that is interacting with an “environment” (here defined as everything that is not the agent). The agent performs actions on the environment (for instance, a robot takes a step forward). Then, the environments will then provide some sort of feedback to the agent, usually in a form called the “reward.”

The agent’s ultimate goal is to maximize the reward. Critically, no one is telling the agent how to maximize the reward or explaining why it gets a reward. the agent needs to figures out how to take actions and observe its environment by itself.

**The introduction of Neural network**

The artificial neural network is kind of machine learning model. It mimics the physiology and functioning of our biological neuron. The basic unit of artificial neural network is neuron. It receives the input from other neuron or external sources and computes an output. Each input () is multiplied by a weight (), which is associate with its relative importance to other input. Next, the neuron is applied an activation function ( to the sum of its weighted input and a bias (). The purpose of the activation function is to introduce non-linearity into the output of a neuron.

****

Now we can combine many neurons into a complicated neural network, also called multi-layer perceptron. The neural network can consist of an input layer, hidden layers, and an output layer. The input layer provides information from the outside world to the network. No computation is performed in any of the Input neurons – they just pass on the information to the hidden neurons. The hidden layer is responsible for computations and transfer information from the input neurons to the output neurons. The output layer also performs computations and transfer the information from the neural network to the outside. Like this neural network, the information moves in only one direction from the input layer, through the hidden layers to the output layer. There are no cycles or loops in the network. We call feedforward network. Later, I will introduce another neural network, called recurrent neural network where the connections between the nodes form a cycle.

****

**Machine learning regression model and property predictor**

Recent years machine learning has been regarded as a burgeoning method for the application to molecular design. Machine learning model has been used as property predictor aiming to predict molecular properties from its structure by learning implicit relation between them. The model could accurately predict thousands of molecules within minutes. Previous works have successfully integrated this technology into molecule screening pipelines[8, 9]. In machine learning framework, molecular structures are transformed into digital representation that serve as input for machine learning model. Two ideal attributes for molecular representation are uniqueness and invertibility. Uniqueness means that a specific molecule can only be represented by a unique molecular representation. Invertibility means molecular representation can be transformed back to a specific single molecule. Actually, not every representation has both two properties. There are mainly two types of representation, including 3D geometry and 2D molecular graphs. The latter can be further subdivided into four categories: string-based, image-based, tensor, and others. In the following, we will briefly describe the Extended Connectivity Fingerprint (ECFP)[10], Simplified molecular-input line-entry system (SMILES)[11], and coulombic matrix[8] methods.

ECFP is a topological fingerprint which originally is used for analysis of molecular characterization, substructure and similarity. Nowadays, it is also adopted to perform machine learning and statistical analysis of molecules. It basically has four steps: 1. Assign each atom with an identifier 2. Atom identifiers are augmented with information from the atom neighborhood 3. Delete the duplicated substructure 4. Hash list of identifiers into a fixed-length bit vector. It does not have the attribute of invertibility. Divided molecular substructure cannot be recombined together into original molecule exactly. Furthermore, due to the fixed length bit vector, each bit may represent multiple substructures, which often leads to the difficulty on analysis. It is worth mentioning that initially each binary bit represents whether the substructure exists in a molecule or not, but herein this study the bit is replaced with the number of the substructure in molecule, as so-called ECFPNUM. ECFPNUM is further used for molecular representation since it is more suitable for property prediction than ECFP.

SMILES is a molecular representation in the form of a line notation for describing the molecular structure by using ASCll characters. For example, a benzene is denoted in the form of SMILES as c1ccccc1. It is the most popular representation in the field of machine learning~~,~~ since it follows particular grammar syntax and can be directly applied into natural language processing (NLP) models. In practice, because many machine learning algorithms cannot process characters(strings) as input directly, it has to be converted into numeric form. The way to standardize converting SMILES into numeric form is first setting every different character as an atom-type. Then for a given molecule every character of its SMILES is converted into bit-vectors formed by the atom-types. The bit-vectors then combined follow the sequence of characters appearing in its SMILES to form a binary matrix which can directly operate as machine learning input. This scheme is known as one-hot encoding. Invertibility is a main advantage of SMILES, since one-hot encoding representation can be converted back to original molecules directly. SMILES, however, also suffers drawback at the same time. One molecule can have multiple SMILES representations. The non-uniqueness SMILES stem from the arbitrary starting atom in a molecule can be used to construct its SMILES. Some cheminformatic packages, such as RDKit[12], have the function to canonize the SMILES. However, Bjerrum et al argues that the latent space created from canonical SMILES may have problem, since only specific grammar syntax has been learned, instead of general underlying rule of molecule structure[13].

Coulombic matrix (CM) is a molecular representation to describe the electrostatic interaction between atoms in a molecule. It is calculated by the equation (1).

(1)

The diagonal element is the polynomial fit of the potential energy of atom itself (self-energy), while the off-diagonal elements correspond to the energy of coulombic interaction between pairs of different atoms in the molecules. The CM can be obtained by performing quantum mechanical calculations. Through the machine learning correlation between molecular structure and its CM could be revealed. Thus, the learned model has exhibited strong ability to predict molecular electronic properties. CM also suffers from a number of issues. Different number of atoms in a molecule lead to different size of coulombic matrix. The solution is to pad by the vacancy of matrix of little molecule. Another issue is that different atom labeling scheme lead to different CM. A simple solution to this is to simply sort the matrices in the order of specific atomic property.

**Property predictor**

Figure 2. 2 shows the prediction performance of three different property-predictor models. The data was collected from Material Project[3] which contains 21165 molecular structures and their corresponding values of IE and EA in the solvent state calculated through *ab* inito calculation based Q-chem software. In this study, 80% of database was used for training, 10% of them was used for validation, and the remaining 10% was used for testing. The first model used ECFPNUM as molecular representation and adopted fully connected neural network (NN) to do machine learning. The second model used SMILES as molecular representation, which is converted into binary matrix through one-hot encoding and adopted recurrent neural network (RNN) to do machine learning. We call this model SMILES(RNN). The last model also used SMILES as molecular representation with one-hot encoding to convert them to digit matrix and adopted convolution neural network (CNN) jointing with a fully connected neural network (NN) to perform machine learning. We call this model SMILES(CNN). The property predictor performance is evaluated by mean absolute error (MAE) on the test set. The results show that using SMILES representation with RNN model can outperform ECFPNUM representation with NN model and the SMILES representation with CNN jointing NN model. Although these models can predict IE/EA from molecular representation swiftly, the optimal MAE is still about 0.2 eV for EA and 0.15 eV for IE. These errors are still large if one wants to do molecular selection which need the errors smaller than 0.1 eV or less. Thus, the models just provide us a tool to do the first screening of molecules from large-scale candidates. The more exact value should rely on more expensive verification method, such as first-principle calculations or experimental measurements.

一張含有 監視器, 室內 的圖片

自動產生的描述

Figure 2. 2 The prediction performance of three different property predictors

**Inverse design and Deep generative machine learning model**

In this work, several aforementioned strategies for molecule selection have been utilized, which indeed significantly accelerate the pace of material development. In conventional material design. The properties of specific material or molecular structure could be obtained through experiments or simulations. But the process is not quietly efficient as mentioned above since exploring chemical space at large scale is a daunting task. Another strategy has been proposed – starting with desired molecular properties, and inversely searching for ideal molecular structures[14, 15]. The process is known as inverse design. Inverse design can be approached in many different methods. One of these methods is high-throughput virtual screening (HTVS)[16, 17]. HTVS allow developers or researchers rapidly screen millions of molecules and enable them to identify promising molecules. Candidate molecules will be further evaluated by more expensive methodologies. HTVS methodology has been successful at designing high-performing molecules. For example, in organic photovoltaics, molecules have been screened for their HOMO/LUMO energies and photovoltaic conversion efficiency[18, 19]. For organic light-emitting diodes, molecules have been screened for their singlet-triplet gap and photoluminescent emission[20].

Another approach is harnessing the power of deep generative model. The goal of generative model is learning how to generate data whose distribution is like the original data set. Generative model can be successfully applied in many different domains, such as image regeneration, audio regeneration and even drug discovery and molecule regeneration. There have been three main generative models mainly used in molecular generation: Variational autoencoder (VAE)[21, 22], Generative adversarial network (GAN)[23, 24], and reinforcement learning(RL)[25, 26]. Here, we will not describe the details of the underlying mechanism of each model, but just briefly introduce them. VAE is derived from autoencoder (AE) shown as Figure 2. 3(b). AE consists of an encoder and a decoder. Encoder is responsible for compressing molecular representation into a vector in low-dimensional space, known as latent space. Decoder aims for transform a vector back to original molecular representation. Nevertheless, because the sparse of latent space created by AE make it difficult to exploit the space, a probability distribution has been introduced by VAE in order to fill uncovered latent space. Therefore, a molecule is not mapping as a fixed point in the latent space, but as a probability distribution, usually Gaussian distribution. The continuous representation of molecules in latent space allow us to generate molecule easily by simple operation, such as decoding random sampling vector, perturbing known chemical structure, or interpolating between molecules. Another generative model is GAN. In this frame work, generator competes against discriminator. Generator attempt to generate artificial data from sampling noise space. Discriminator is responsible for determining whether the data is fake or real. The goal of this model is that generator has ability to generate data which discriminator cannot distinguish as fake. As a result, we can generate new molecules from the generator. The last one is reinforcement learning (RL), which has been seen as the most promising model in molecular design recently. The mechanism of RL is that the generator learns how to take action in order to maximize the reward (properties) within an environment (SMILES generation).

Figure 2. 3 Graphical diagram of machine learning application to molecular design. The figure was adapted from Kang et al[27] (a) property prediction, (b) autoencoder, (c) variational autoencoder proposed by Bombarelli et al[21], (d) variational condition autoencoder proposed by Kang et al[27]



**Data**

There are only few online accessible databases containing EA and IE values, such as Material Project (MP)[3]and National Institute of Standards and Technology (NIST). MP has about 35336 molecular structures and its corresponding IE and EA values (not every molecule labeled both EA and IE) which are acquired through quantum mechanical calculation (continuously updated). Besides, MP provides user a convenient Application Programming Interface (API), so users can easily access the molecular data according to their need. NIST also stores molecular EA and IE values which are acquired by experiments. As NIST do not provide any API, it is hard for users to collect large number of data, and thus web crawling must be involved in automatic collection workflow. Additionally, QM9[4, 8] includes 133885 molecular SMILES and its corresponding Highest Occupied Molecular Orbital (HOMO) and Lowest Unoccupied Molecular Orbital (LUMO). The HOMO and LUMO can be roughly mapped to EA and IE, since there is high correlation between IE/HOMO and EA/LUMO. In terms of dipole moment and stability of molecule structure, as we know, there is a dipole moment database in NIST. At the same time, we are looking for chemical properties which can quantitative describe the stability of molecule. Furthermore, in order to train deep generative model, collecting a large number of molecules is necessary. In our work, six million molecular SMILES data from ZINC[5] is collected to train the deep generative model. ZINC is a free database of commercially-available compounds and contains over 750 million molecules that can be downloaded efficiently.

**Our adapted model and experience**

Since VAE has been introduced into molecular design in 2016 by Bombarelli et al, over forty works has been proposed by using different deep generative model and some variation of them. In this work, to design ideal molecules for electrolyte, two proposed models have been adapted to our generative model for designing electrolyte molecules. The first one is chemical VAE (chemvae), proposed by Bombarelli et al[21](Figure 2. 3(c)), and the other one, named conditional-molecular-design-ssvae (SSVAE), is proposed by Kang et al[27](Figure 2. 3(d)). In the frame of chemvae(Figure 2. 4), SMILES is transformed into a one-hot representation (**x)**, and convolution neural network (CNN) is used as encoder to map **x** to latent space, **z**, with Gaussian distribution. The RNN decoder is to transform **z** back to one-hot representation, **x**. Additionally, another neural network, named property predictor, is mapping **z** to molecular properties. shows the structure of the model. The continuous representation of molecules in latent space allows us to generate molecule easily by simple operation, such as decoding random sampling vector, perturbing known chemical structure, or interpolating between molecules. This work also introduces the Gaussian process model to guide the search of ideal molecules efficiently in latent space. In Bombarelli’s model three components (encoder, decoder, and property predictor) were trained simultaneously. In this study, a small portion of the data used in this work is obtained from MP with IE/EA label and many other unlabeled molecules whose SMILES are from ZINC, hence the VAE and properties predictor cannot be trained simultaneously. A new approach is that VAE model can be trained by all the molecular SMILES from MP and ZINC first. After finishing VAE training, the encoders from VAE are extracted and combined with property predictor neural network for the next step of training MP data containing molecular SMILES and its corresponding properties. In order to preserve VAE system, in the process of training this property predictor, all the parameters in the encoder need to be freezed and only optimize the parameters of property predictor. However, the prediction performance of this model is not good enough to compare with our previous trained property predictors, since the parameters of encoder cannot be optimized accordingly. In Bombarelli’s original work, all the training data are labeled, but in our case the labeled data only counts a small portion of training data. The paper shows that latent space created by autoencoders jointly trained with the property prediction task shows a gradient by property values; molecules with high specific property values are located in one region,



Figure 2. 4 The structure of chemvae

and molecules with the low values are in another. Autoencoders that were trained without the property prediction task do not show a discernible pattern with respect to property values. It seems like encoder, decoder and property predictor if trained together molecules mapped in the latent space show property related distribution. Evidently, the chemvae model is not suitable for our case, since the molecules labeled with IE/EA only counts a small portion of all data in our case.

Figure 2. 5 The structure of SSVAE

In reality, small portion of molecules were labeled with their properties is a very common situation in most database because the cost of acquiring properties is very expensive. Kang et al try to solve this problem and proposed conditional-molecular-design-ssvae (SSVAE)[27]. The model combined semi-supervised learning and conditional autoencoder together and allows us to directly generate new molecules whose properties are close to a predetermined target condition without any extra optimization procedure. In this frame work, the labeled data are used for building a property predictor which introduces the gaussian distribution in order to address the intractability of **y**. The missing **y** of unlabeled data are predicted by this property predictor. Next, the RNN encoder maps **x** and **y** to the latent space **z,** similarly introducing the gaussian distribution. Lastly, RNN decoder maps **y** and **z** to original **x**. The structure of SSVAE is shown in Figure 2. 5. In this work, MP data with IE/EA label and a large number of molecular SMILES from ZINC are used to train the SSVAE model.

The performance of SSVAE is evaluated through our previous trained SMILES(RNN) property predictor model. Based on setting a predetermined IE value, 1000 molecules from the decoder of SSVAE mode are generated then, screened by our previous mentioned SMILES(RNN) property predictor. Figure 2. 6 demonstrates that the value of IE distribution of all the original data and the generated molecules whose IE value were predicted by property predictor. Although the result is not as good as the paper reported [27], the trend of distribution is consistent with the increase of IE values. In this study, IE value is set to ~~4~~, 5 and 6 eV. The tolerance for the screening is 0.2 eV. The success rates are 15.3%, 22.9%, and 15.2% for 4, 5, and 6 eV, respectively.

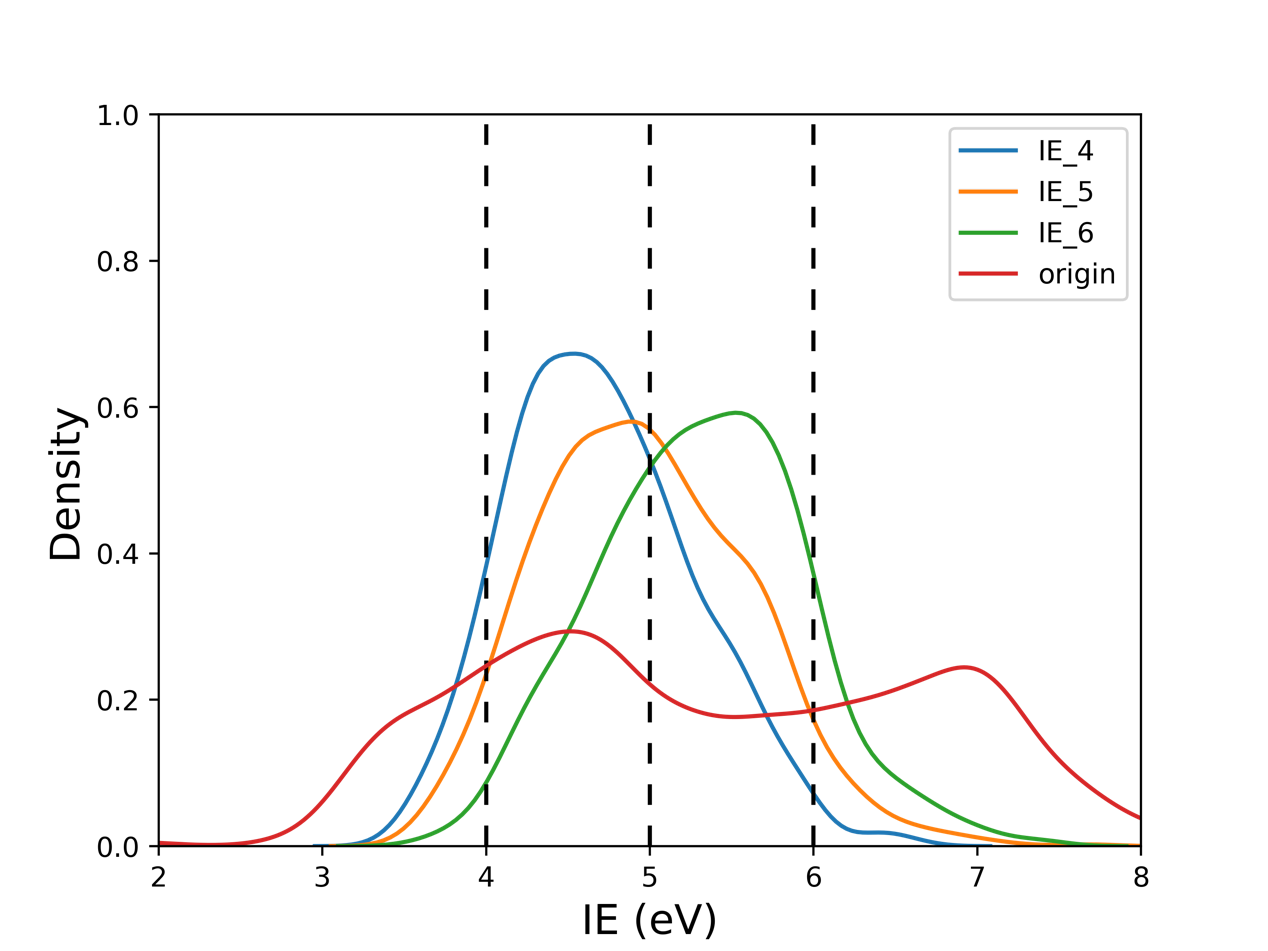


Figure 2. 6 The IE distribution obtained by the generative model SSVAE

**The determination of IE/EA**

Candidate molecules can be synthesized and characterized by various techniques for electrochemical testing. It is worthwhile noting that the energy level of IE can be measured by ultraviolet photoelectron spectroscopy and inverse photoemission spectroscopy[28]. The main principle of photoelectron spectroscopy is that the source of radiation can eject molecular electrons from core orbitals or valence orbitals. The ionization energy is obtained by subtracting the energy of the source radiation from the measured kinetic energy of the ejected electron. For the energy level of EA, it can be also measured by photoelectron spectroscopy and laser photodetachment[29]. Both methods directly measure a threshold energy for the removal of an electron from the molecular anion. In addition, by using voltammetric technique, the redox potentials (reduction/oxidation potentials which are respectively proportional to IE/EA values) can be verified by cyclically sweeping a potential between a working electrode and a reference electrode in the solution, while measuring the current response[30]. The experimental results can be compared with the density functional theory calculations results.

**Compared experiment and first-principles calculation**

Comparing conventional electrolyte molecules – EC, DMC, and VC, their IE values are obtained from experiments (NIST database) and first-principles calculation (Material Project database). The experimental values of IE are 10.4, 11.0, and 10.08 eV for EC, DMC, and VC respectively[31, 32]. And the first-principles calculation values of IE are 8.44, 8.3, and 7.04 eV for EC, DMC, and VC respectively[3]. The order of experimental IE value of EC and DMC is not consistent with the order of first-principles calculation value. The discrepancy in order can be probably attributed to a several of factors. First, the experiments measure the value of IE in the gas phase, but for the first principles calculation in MP, solvent environment has been taken into consideration. Furthermore, first-principles calculation solves the Schrodinger equation through series of approximations and simplifications. The final result could vary by the selection of different basis sets and DFT functionals. Nevertheless, in both methods, the IE value of DMC and EC are higher than that of VC, which means VC could be oxidized more easily than DMC and EC. The result suggests that VC can be a suitable electrolyte additive molecule, and EC and DEC can be suitable for electrolyte molecules.

**Conclusions**

In this work, the power of advanced machine learning technology has been harnessed to design brand new electrolyte molecules which are suitable for high-voltage lithium-ion battery electrolyte. The energy level of IE/EA of molecules is regarded as a metric to gauge the potential of candidate molecule. Three property predictors which have ability to map molecular structure to its chemical properties (IE/EA) rapidly have been built. Among them, SMILES (one-hot encoding) with RNN has better prediction performance than the others. Furthermore, in order to carry out inverse design, the proposed generative model is adapted to our application. So far, new molecules with predetermined property values can be automatically generated. Nevertheless, unlike the other fields such as particular in speech recognition and computer vision, the scarcity of labeled data in material and molecular database significantly impedes the development of machine learning in this field. In addition, even though machine learning exhibits phenomenal prediction ability, the difficulty of interpreting machine learning models makes people hard to unveil the real latent relation between the molecule and its corresponding properties. Undoubtedly, machine learning in molecules and material design deserve more material or molecular knowledge involved. All in all, this work is expected to pave the way for resolutions to the aforementioned problems and electrolyte molecular design.

**References**

1. Sun, Y., N. Liu, and Y. Cui, *Promises and challenges of nanomaterials for lithium-based rechargeable batteries.* Nature Energy, 2016. **1**(7): p. 16071.

2. Fukui, K., T. Yonezawa, and H. Shingu, *A Molecular Orbital Theory of Reactivity in Aromatic Hydrocarbons.* J. Chem. Phys., 2004. **20**: p. 722-725.

3. Qu, X., et al., *The Electrolyte Genome project: A big data approach in battery materials discovery.* Computational Materials Science, 2015. **103**: p. 56-67.

4. Blum, L.C. and J.-L. Reymond, *970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13.* J. Am. Chem. Soc., 2009. **131**(25): p. 8732-8733.

5. and, J.J.I. and B.K. Shoichet\*, *ZINC − A Free Database of Commercially Available Compounds for Virtual Screening.* J. Chem. Inf. Model., 2004. **45**(1): p. 177-182.

6. Maine, E. and E. Garnsey, *Commercializing generic technology: The case of advanced materials ventures.* Res. Policy, 2006. **35**(3): p. 375-393.

7. Virshup, A.M., et al., *Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds.* J Am Chem Soc, 2013. **135**(19): p. 7296-303.

8. Rupp, M., et al., *Fast and accurate modeling of molecular atomization energies with machine learning.* Phys Rev Lett, 2012. **108**(5): p. 058301.

9. Hansen, K., et al., *Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space.* J Phys Chem Lett, 2015. **6**(12): p. 2326-31.

10. Rogers, D. and M. Hahn, *Extended-Connectivity Fingerprints.* 2010.

11. Weininger, D., *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules.* J. Chem. Inf. Comput. Sci., 2002. **28**(1): p. 31-36.

12. Landrum, G., *RDKit: Open-source cheminformatics.*

13. Sattarov., E.J.B.a.B., *Improving Chemical Au- toencoder Latent Space and Molecular De novo Generation Diversity with Heteroencoders.* arXiv e-prints:1806.09300, June 2018.

14. and, C.K. and David N. Beratan\*, *Inverse Strategies for Molecular Design.* J. Phys. Chem., 1996. **100**(25): p. 10595-10599.

15. Zunger, A., *Inverse design in search of materials with target functionalities.* Nature Reviews Chemistry, 2018. **2**(4): p. 1-16.

16. Broach, J. and J. Thorner, *High-throughput screening for drug discovery.* Nature, 1996. **384**(6604): p. 14-16.

17. *Discovery of small molecule cancer drugs: Successes, challenges and opportunities - Hoelder - 2012 - Molecular Oncology - Wiley Online Library.* 2019.

18. Xiao, D., et al., *Inverse Design and Synthesis of acac-Coumarin Anchors for Robust TiO2 Sensitization.* J. Am. Chem. Soc., 2011. **133**(23): p. 9014-9022.

19. S. A. Lopez, B.S.-L., J. de Goes Soares, A. Aspuru-Guzik, *Design principles and top non-fullerene acceptor candidates for organic photovoltaics.* Joule, 2017. **1**(4): p. 857–870.

20. Gómez-Bombarelli, R., et al., *Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach.* Nature Materials, 2016. **15**(10): p. 1120-1127.

21. Gómez-Bombarelli, R., et al., *Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules.* ACS Central Science, 2018. **4**(2): p. 268-276.

22. Kingma, D.P. and M. Welling, *Auto-Encoding Variational Bayes.* 2013.

23. Goodfellow, I.J., et al., *Generative Adversarial Networks.* 2014.

24. Kadurin, A., et al., *druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico.* Mol Pharm, 2017. **14**(9): p. 3098-3104.

25. Olivecrona, M., et al., *Molecular de-novo design through deep reinforcement learning.* Journal of Cheminformatics, 2017. **9**(1): p. 1-14.

26. Popova, M., O. Isayev, and A. Tropsha, *Deep reinforcement learning for de novo drug design.* Science Advances, 2018. **4**(7): p. 7885.

27. Kang, S. and K. Cho, *Conditional Molecular Design with Deep Generative Models.* Journal of Chemical Information and Modeling, 2018. **59**(1): p. 43-52.

28. Akaike, K., et al., *Ultraviolet photoelectron spectroscopy and inverse photoemission spectroscopy of [6, 6]-phenyl-C 61-butyric acid methyl ester in gas and solid phases.* Journal of Applied Physics, 2008. **104**(2): p. 023710.

29. Kebarle, P. and S. Chowdhury, *Electron affinities and electron-transfer reactions.* Chemical Reviews, 1987. **87**(3): p. 513-534.

30. Brushett, F.R., J.T. Vaughey, and A.N. Jansen, *An all‐organic non‐aqueous lithium‐ion redox flow battery.* Advanced Energy Materials, 2012. **2**(11): p. 1390-1396.

31. McGlynn, S. and J. Meeks, *Photoelectron spectra of carbonyls, carbonates, oxalates and esterification effects.* Journal of Electron Spectroscopy and Related Phenomena, 1976. **8**(2): p. 85-93.

32. Wittel, K., et al., *Photoelectron Spectra and Molecular Properties, Xlviii Carbonates and Thiocarbonates.* Zeitschrift für Naturforschung B, 1975. **30**(11-12): p. 862-874.