Prediction and design of new molecules with electrolyte chemical properties via Variational Autoencoder

Cheng-Yang Tu (杜政陽)(1), Jia-Han Li (李佳翰)(1), Wen-Jay Lee (李玟頡)(2)

(1)Department of Engineering Science and Ocean Engineering, National Taiwan University , No. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan, 10617

(2) National Center for High Performance Computing, No. 22, Keyuan Rd., Central Taiwan Science Park, Taichung, Taiwan, 40763

Abstract

Traditionally, exploring the new materials and their properties requires to repeat a number of experiments for getting their chemical properties. Although the current methodologies such as high-throughput computation[1] has been developed for the novel materials screening and exploration, they are still costly and time consuming. Therefore, computational materials experts have introduced machine learning to accelerate the discovery of new materials recently. In this work, the electrolyte additive is selected as an example to demonstrate the effectiveness of this model. Variational autoencoder (VAE)[2] and recurrent neural networks(RNN) were employed to assist the design of chemical structure for specific applications, which efficiently generate the candidates with desired properties.

Problem description

In recent years, lithium-ion batteries (LIB) have been widely used in electric vehicles, consumer electronics, and stationary energy storage systems. Improving safety and high energy density/performance is the goal pursued by LIB, so the primary goal of improving performance is to change its electrolyte. Unfortunately, the development of new materials has always been quite time-consuming and cost-consuming, but deep learning (DL) provides a potential solution that can shorten the development cycle and explore new chemical structures. It is not necessary to repeatedly experiment to obtain new molecules through simulations and experiments. Instead, a large number of novel candidate molecules with predictive properties can be obtained through the DL model. Since it is difficult to search a large area of chemical space, we need to build a general model to search for the best molecule described by continuous representation. In addition, deep learning methods are used to construct a polymer electrolyte generation model to facilitate the future design of new polymer electrolytes or materials.

Method

The design of a suitable electrolyte additives for a stable SEI layer is crucial and difficult in LIB, because they should satisfies the specific conditions of electrochemical properties. In this work, deep learning(DL) was used to assist the design of electrolyte additives as shown in workflow of figure 1. The dataset of common additives(111 molecules), QM 9(133274 molecules), Wohlfarth et al. [4] (784 molecules), and Martin et al. [5] (5445 molecules) are selected as the DL datasets, where they included electrochemical properties of HOMO, LUMO, bandgap, dipole moment, Dielectric constant (DC), ionization energy (IP), and electron affinity (EA). In figure 1, variational autoencoder as a generate model was trained with QM9 dataset, which was employed to produce the new electrolyte additives with desired properties. PCA analysis was used to characterize the latent space for visualizing chemical properties and selected the desired property range. In the following, RNN models were trained with datasets of Wohlfarth et al. [4] and Martin et al. [5] for prediction of DC, IP, and EA. In order to make sure that the data were close to data of QM9, the data were projected on the latent space trained with QM9 as shown in figure 4. The final candidates were determined with the screened condition after the prediction of RNN model as shown in figure 1.

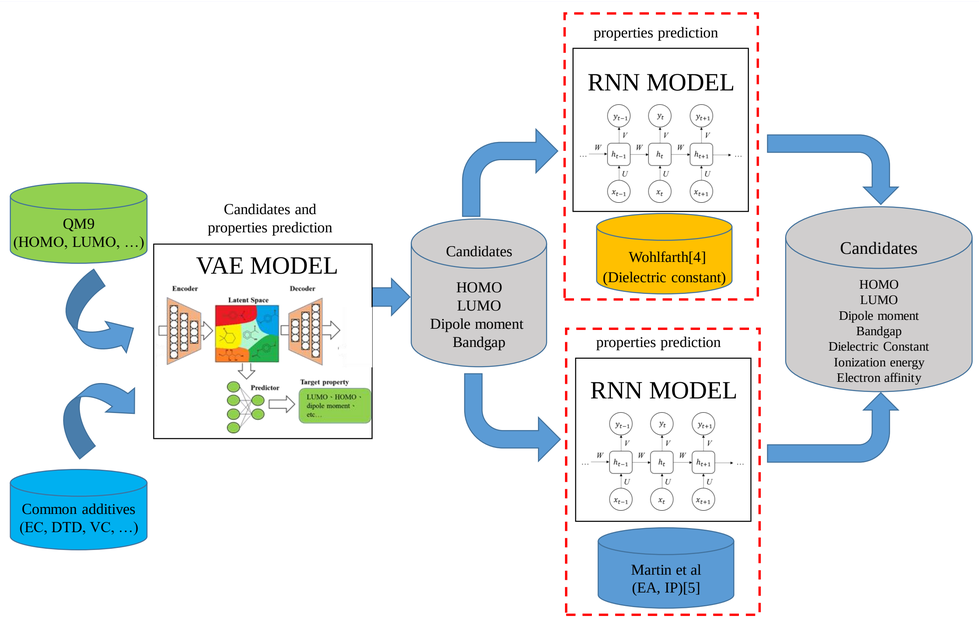

Figure 1 Workflow for selecting new molecule candidates.

Results

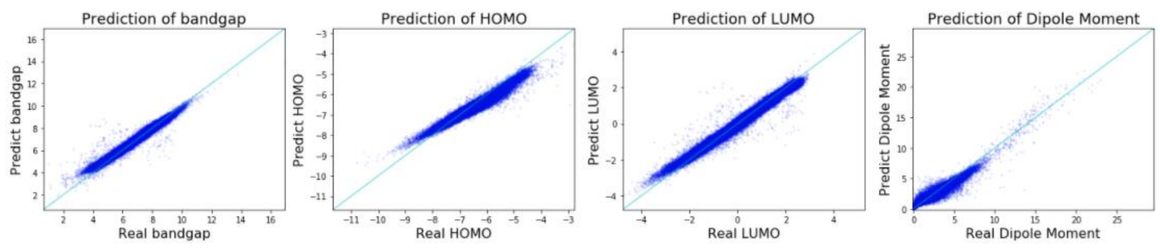

Figure 2 Relation of chemical properties between the real value of properties from the QM 9 validation dataset and predicted properties from the VAE model.

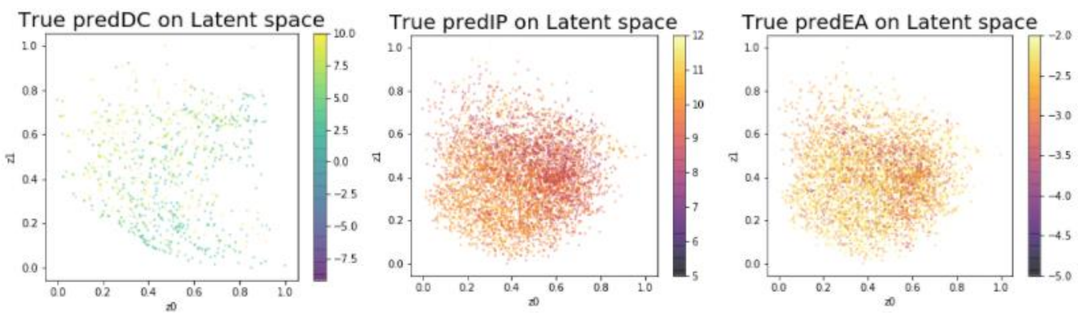

Figure 4 Mapping the datasets of Wohlfarth et al. [4] and Martin et al. [5] on the latent space of VAE.

Table 1 property prediction results of QM9 and the other dataset.

	bandgap	HOMO	LUMO	Dipole moment	Dielectric Constant	IP	EA
PearsonR	0.9821	0.9502	0.9876	0.9430	0.8930	0.8737	0.8101
MAE	0.1962	0.1570	0.1978	0.4153	0.3948	0.3678	0.3487
R2	0.9589	0.8806	0.9601	0.8701	0.7805	0.7609	0.6512

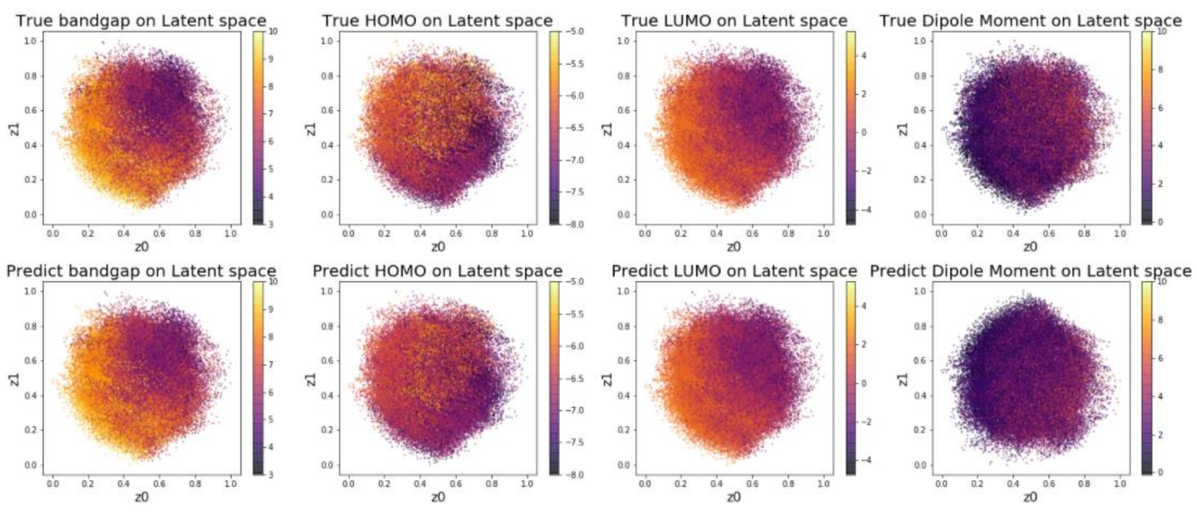

Figure 3 The PCA analysis of the latent space, the selected properties in the QM9 data are presented in a two-dimensional mode with their respective PCA analysis.

Table 2 The 20 potential chemical structures determined from the screening workflow.

smiles	bandgap	HOMO	LUMO	Dipole Moment	Dielectric constant	IP	EA	chemical hardness
O=CC(=O)C=O	3.2367	-6.972	-3.7017	3.1202	38.312	11.5935	0.6618	5.46585
O=CC(=O)C(=O)CO	3.8061	-6.8563	-3.0099	3.1251	30.3472	11.054	0.6002	5.2269
O=C1COCC(=O)C1=O	3.8751	-6.9164	-2.9901	3.5944	42.7701	10.575	0.7396	4.9177
[NH-]c1nc[cH+]nc(O)n1	4.0476	-6.9718	-2.8409	5.0148	36.3849	7.0077	-2.702	4.85485
Nc1nonnc1=O	3.9978	-6.8631	-2.8123	3.5888	48.557	12.6697	0.7279	5.9709
NC(=O)C(=O)C=O	3.7784	-6.6052	-2.7735	3.0645	40.2403	11.8793	-0.2964	6.08785
O=C1[CH-]C2=CC=CC2N1	3.2125	-6.0226	-2.7706	7.3101	44.422	9.4655	-2.7355	6.1005
COC(=O)C(=O)C=O	4.0775	-6.8429	-2.7313	3.0398	38.0667	11.1648	-0.1542	5.6595
[NH-]C1N=CC=NO1	4.1451	-6.9024	-2.726	3.044	38.5198	8.6224	-2.2891	5.45575
O=C1CCCC(=O)C1=O	4.1071	-6.8899	-2.7012	4.8377	42.0834	10.3533	0.6891	4.8321
O=CC(=O)C(=O)N1CC1	4.2132	-6.9402	-2.695	3.2035	42.5466	10.0775	-0.6088	5.34315
N=C(N)C(=O)C=O	4.0458	-6.8098	-2.6841	4.0772	14.6068	11.448	-0.8138	6.1309
[NH-][CH+]OCC(=O)C=O	4.1929	-6.9033	-2.6804	3.1177	19.5452	9.7328	-1.4855	5.60915
[NH-]C1=NC=C[CH+](C=O)O1	4.1445	-6.8548	-2.6769	3.0552	39.212	9.1977	-2.1783	5.688
CNC(=O)C(=O)C=O	3.8836	-6.6138	-2.6726	3.2299	95.1192	11.6288	-0.7955	6.21215
O=C1CNCC(=O)C1=O	3.9592	-6.6913	-2.6668	3.6193	44.3411	10.7982	-0.2631	5.53065
CNC(=N)C(=O)C#N	4.0393	-6.7318	-2.6401	3.3159	77.8812	11.829	-0.6964	6.2627
O=C1COCC1=O	3.7943	-6.4717	-2.6153	3.12	49.3587	10.3211	-0.4988	5.40995
O=CC1=CC2OC2C1=O	4.2104	-6.8683	-2.607	3.5894	10.3363	10.5864	-0.8819	5.73415
Nc1cnc(=O)on1	4.3241	-6.9613	-2.5978	6.1211	48.9019	10.5188	-1.9318	6.2253

Conclusions

To find the features of the target molecule, another feature predictor is built in the latent space. The predictor is an ANN architecture. VAE is trained by 133274 molecules in the QM 9 dataset [3]. The generated model has trained on QM 9 dataset[3] which contains different databases are related, the model is generated to assist in the design of chemical structure HOMO, LUMO, bandgap, and dipole moment. In addition, we also use two different databases with 784 features of Dielectric constant (DC) and 5445 features of ionization energy (IP) and electron affinity (EA). We also set up the RNN model to combine these two data The library is trained through RNN, and the trained molecules through the VAE model, through PCA analysis, to check the versatility of QM9 and various databases. Finally, we searched 20 potential molecules as candidate in the latent space with the chemical properties that can stabilize LIB, and Table 2 shows the various properties of the candidates.

References

[1] Halls, M. D et.al(2010). 1472-1478..

[2] Yan, C. et.al(2020) (pp. 1-7).

[3] Wu, Z. et.al (2018) 513-530.

[4] Bouteloup, R. et.al. (2019) 11043-11057..

[5] Korth, M. et.al (2014) 7919-7926.