Computational Materials Science 228 (2023) 112361

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

Full Length Article

Transfer learning for predicting reorganization energy

Xushi Zhang [a], Guodong Ye [b,*], Chuanxue Wen [a,*], Zhisheng Bi [a,*]

[a] School of Biomedical Engineering, Guangzhou Medical University, Guangdong 511436, China
[b] School of Pharmaceutical Sciences, Guangzhou Medical University, Guangdong 511436, China

ARTICLE INFO

Keywords:
Transfer learning
Reorganization energy
Frontier molecular orbital energy

ABSTRACT

Reorganization energy, a crucial factor in quantum chemistry, plays an integral role in the research of optoelectronic and electrical devices such as organic field-effect transistors and organic light-emitting diodes. Currently, the calculation of reorganization energy predominantly depends on density functional theory, which poses significant computational costs. Traditional machine learning, commonly employed to predict reorganization energy necessitates a substantial volume of data for model training, a process that can be time-consuming due to difficulties in collecting sufficient data. This study proposes a novel solution by employing transfer learning. This method involves the pre-trained model using publicly accessible large-scale data on frontier molecular orbital energy, and then transfers this model to predict the reorganization energy. The experimental results indicate that the proposed method requires approximately 28% less data compared to traditional machine learning while maintaining equivalent accuracy. Moreover, this method achieved a median error of less than 0.02 eV when predicting low reorganization energy. Consequently, the need for reorganization energy data in model construction is significantly reduced, thus facilitating research into reorganization energy and expediting research into the optical properties of materials.

1. Introduction

With affordability, compatibility for large-area fabrication, and low power consumption [1], organic semiconductors are imperative for solar energy conversion and various optoelectronic devices [2–4] including light-emitting diodes [5], organic solar cells [6–8], field-effect transistors [9,10], chemical sensors [11–15], and more [16]. Charge mobility, critical characteristics of semiconducting devices [17], provides significant insights into their physical properties. Molecules with low reorganization energies are crucial in developing systems with efficient charge transport, facilitating an acceleration of the intermolecular charge hopping rate [18–21] and improving the emission quantum yield for intramolecular excitations by reducing the non-radiative decay rate [22,23].

The reorganization energy (RE) describes the total energy change associated with the deformation of the lattice and molecular structure as the charge transfers [24]. Theoretically, RE comprises a major internal component and a minor external one [25–27]. The intramolecular RE is used to measure the energy shift in a system caused by the relaxation of its geometric structure, whereas the latter corresponds to the energy necessary to repolarize surrounding environmental molecules. A key component of the Marcus theory [28–31] states that the internal RE, which articulates the energy change required to distort geometry upon charge transfer, is an essential charge transport parameter suitable for molecular-level screening. Moreover, the internal RE plays a vital role in determining the charge transfer rate and resultant charge mobility [32–34]. RE, denoted as $\lambda(h)$, is calculated using Eq. (1). In this equation, $\lambda(I)$ represents the vertical energy change caused by structural relaxation from the unstable point to the stable point in the excited state, while $\lambda(II)$ represents the relaxation energy in the ground state as shown in Fig. 1. The traditional method for calculating RE relies on density functional theory (DFT) calculations, requiring two geometry optimizations. This process necessitates significant computational resources and can take anywhere from hours to weeks to complete, contingent on the system size and calculation type. Therefore, machine learning (ML) has been proposed as a method for predicting RE, with it offering considerably faster processing times. For example, Abarbanel et al. applied ML to predict the intramolecular RE of a wide range of polythiophenes [35]. Moreover, ML are significantly faster than semi-empirical computations such as PM7, even though these semi-empirical computational methods are faster than DFT calculations [36].

$$\lambda(h) = |\lambda(I) + \lambda(II)| \tag{1}$$

* Corresponding authors.
E-mail addresses: gzhygd@gzhmu.edu.cn (G. Ye), wenchuanxue@gzhmu.edu.cn (C. Wen), bivictor@gmail.com (Z. Bi).

https://doi.org/10.1016/j.commatsci.2023.112361
Received 29 March 2023; Received in revised form 29 June 2023; Accepted 30 June 2023
Available online 5 July 2023

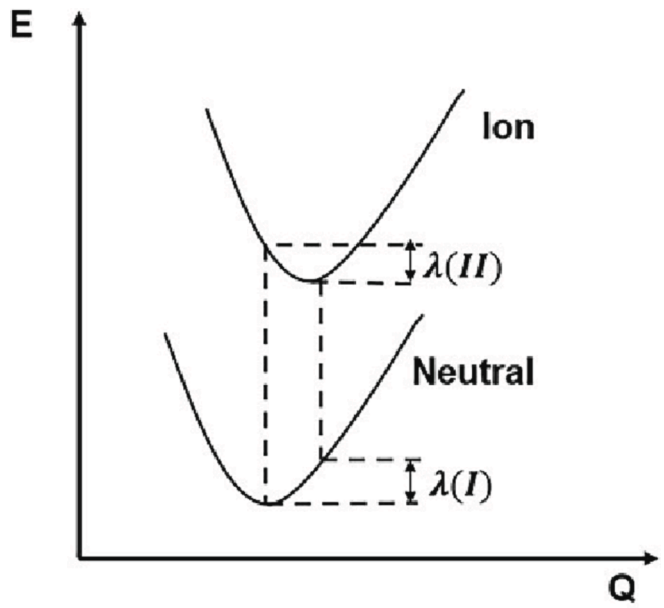

Fig. 1. Schematic diagram of RE.

Recently, ML has garnered substantial interest from the scientific community owing to its remarkable potential to reduce computational costs and expedite atomistic calculations [37–42]. It has been widely implemented in materials research, electronic structure property calculations, and drug discovery [43–51]. Nevertheless, ML necessitates a large volume of datasets to train a network for accurate predictions. Given the complexity and high computational cost of calculating the RE, obtaining an ample RE for network training is challenging, leading to inaccuracies in predictions when the RE data is scarce.

To solve this problem, transfer learning (TL) is utilized to predict the RE when sufficient training data are not available. TL can adapt data from one task and repurpose it as the starting point for various correlated tasks, proving useful when training data are limited [52,53]. Its applications in chemistry are manifold, including the use of models pre-trained on small molecules for larger ones [54], pre-training models for quantitative structure property/activity relationship (QSPR/QSAR) predictions [55,56], and transferring models trained on DFT data to CCSD(T)/CBS data [57].

In the quest for new molecules, calculating the frontier molecular orbital energy (E_FMO), including the highest occupied and lowest unoccupied molecular orbitals (HOMO and LUMO), is vital, particularly for estimating optoelectronic properties and screening databases of potential organic molecules [58]. The HOMO is a molecule's propensity to lose electrons, while the LUMO depicts its ability to accept electrons. The LUMO describes the process wherein a molecule's electron gets excited from an unoccupied to an occupied orbital. The influence of an electric field on the HOMO and LUMO is considered a crucial factor in evaluating the suitability of organic materials as conducting channels in OTFTs [59]. Frontier molecular orbital analysis is leveraged to predict physicochemical and quantum chemical properties and estimate potential intra- and intermolecular interactions as along with the optical properties of all stable conformers of the examined compound [60]. E_FMO correlates with RE [24], suggesting a potential for RE magnitude quantification through molecular orbital properties [61]. E_FMO can inform the OLED design materials with optimal charge-carrier transport performance, and compounds with desirable transport characteristics can be designed by controlling the RE differential [62]. These findings indicate that E_FMO and RE possess the capacity to determine a compound's electron transfer. Previous studies [35,36] also indicate that E_FMO and RE can be calculated based on a compound's structural characteristics, such as molecular fingerprints, suggesting a correlation between E_FMO, RE, and a compound's structure. Therefore, this study proposes utilizing a substantial volume of E_FMO data to pre-train the network and fine-tune it in the TL process for the RE prediction of small samples.

Creating a sample of chemical quantities using traditional calculation methods can be time-consuming. TL mitigates this issue by using chemical quantities calculated via a particular method to pre-train the network, repurposing it to predict different chemical quantities. This strategy considerably reduces the cost and time needed for creating samples used in network training. By pre-training the model with E_FMO data from another dataset, TL can effectively predict the RE. Furthermore, TL can deliver better results even with less quantity of RE data. Subsequently, Section 2 details the composition of training data and the strategy. In Section 3, the performance of TL models pre-trained by the HOMO or LUMO in predicting the RE is evaluated. Moreover, the influence of different molecular fingerprints on the TL is discussed. Finally, Section 4 concludes the paper.

2. Methods

2.1. Dataset

The database consists of 111,725 molecular structures and E_FMO values, calculated using the B3LYP [63] method with the 6-31G* basis set [64], where are deposited in a public repository [65]. The database's molecular structures include the atomic elements C, H, B, N, O, F, Si, P, S, Cl, Se, and Br, primarily originating from organic electronic materials. SDF files, generated using the ChemAxon Standardizer [66] and OpenBabel [67], were applied to the tagged molecule in this database. Owing to a substantial number of similar compounds and repeated E_FMO values in this dataset, we removed the redundant values and randomly selected 75,000 molecules as our final dataset, which we denote as Dataset75000. In this dataset, HOMO energies range from −9.49 to −2.98 eV, and LUMO energies range from −4.35 to 2.70 eV as depicted in Fig. 2. a.

We collected a dataset of 7681 hole RE values from publicly accessible websites [68], labeling it REdataset7681 for differentiation. This dataset consists of dimers, tetramers, and hexamers composed of two, four, and six monomers, respectively, selected from the 253 available thiophene-based monomers. The DFT-calculated RE ranges between 0.01 and 1.96 eV (as depicted in Fig. 2. b). The DFT calculations were performed using the B3LYP [63] functional with the 6-31G* basis set [64]. DFT requires an approximation for the exchange–correlation energy as a function of the electron ground-state density in practical calculations [69]. It can be leveraged to study redox reactions and determine RE by analyzing the gain, loss, and transfer of electrons under an electric field's influence. However, systems containing transition metals impose stringent accuracy requirements on density functional approximations [70]. The E_FMO and RE values considered in this study are derived from organic molecules, primarily composed of nonmetallic elements such as carbon, hydrogen, oxygen, and nitrogen. Therefore, DFT-calculated E_FMO and RE maintain a certain degree of accuracy. For example, the Harvard Clean Energy Project used DFT to screen two million organic compounds containing E_FMO in its research on high-efficiency organic photovoltaic materials [71].

To explore the impact of different RE datasets on TL further, we employed an additional RE dataset in subsequent experiments. This dataset, incorporating electron and hole RE, originated from the chromophore v1 dataset [72] within the OCELOT database [73]. It encompasses DFT or DFT/TDDFT, that is, time-dependent DFT, computed properties for organic, π-conjugated molecules, and crystal structures relevant to electronic and light-oriented technologies. This dataset consists of 25,251 organic π-conjugated, along with their electronic, redox, and optical properties calculated by the high-accuracy DFT/TDDFT method, utilizing IP-tuned LC-ωHPBE functionals and the Def2SVP basis set [74–76]. TDDFT allows for predictions of electronic

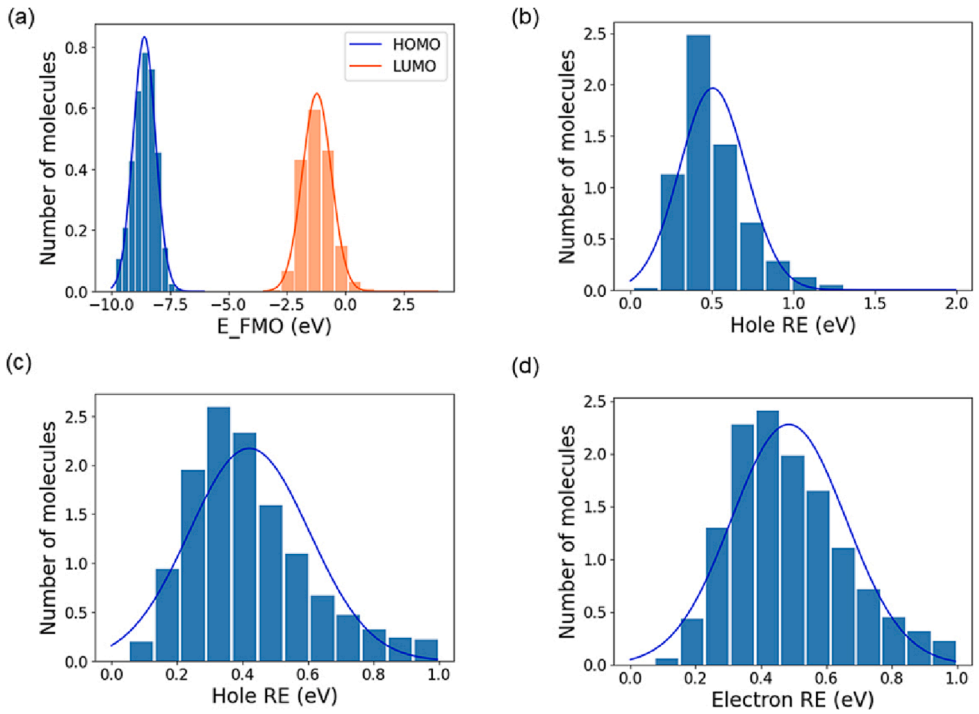

Fig. 2. Histogram of datasets considered in this paper. (a) E_FMO in the Dataset75000, (b) Hole RE in REdataset7681, (c) Hole RE in REdataset25251, and (d) Electron RE in REdataset25251.

excitations in terms of both transition frequency and oscillator strength. These corrections pertain between the occupied and unoccupied levels of the ground-state KS potential [77]. TDDFT is a tool often used in modern theoretical chemistry to study electronic excited states [78]. Since RE is associated with a molecule's excited states, DFT/TDDFT is particularly suitable for calculating the RE. The molecules within this dataset are fragments of experimentally synthesized organic compounds containing elements such as C, N, O, F, S, Cl, Br, Se, P, Si, B, As, Te, I, and H. The dataset contains molecules with varying numbers of π-conjugated rings, ranging from one for benzene derivatives to as many as 28 for large π-conjugated systems, including fullerene derivatives. 60% of the molecules in the dataset, such as biphenyl, do not have fused-aromatic rings, whereas the remaining 40%, such as naphthalene, have at least one fused-aromatic ring. We used the RE values in the 25,251 molecules to establish a second dataset, which we named REdataset25251. The electron RE values in this dataset range from 0.07 to 1.00 eV, and the hole RE values span from 0.05 to 1.00 eV (as demonstrated in Fig. 2.c and Fig. 2. d). Moreover, 16% of the hole RE values fall between 0 and 0.25 eV.

Tanimoto coefficient is a metric to measure the similarity of two molecular fingerprints. The Tanimoto coefficient (T) of molecular fingerprint A and molecular fingerprint B can be defined as Eq. (2). In this equation, a represents the number of 1 in molecular fingerprint A, b is the number of 1 in molecular fingerprint B, while c represents is the number of 1 that are shared in A and B. Since some datasets only have ECFP4 (molecular fingerprints), we analyzed the similarity of compounds through the Tanimoto coefficient of ECFP4 [79], and the results are shown in Fig. 3. It can obtain an average Tanimoto coefficient of 0.07 between Dataset75000 and REdataset7681, and an average Tanimoto coefficient of 0.06 between Dataset75000 and REdataset25251. Analysis found that there are 136 pairs of compounds with a Tanimoto coefficient equal to 1 in REdataset25251 and Dataset75000, suggesting there are 136 pairs of compounds with the same molecular fingerprints. However, ECFPs cannot be reversed into compound structure [80], it cannot be demonstrated that these 136 pairs of compounds are completely consistent. Overall, the average Tanimoto coefficient of the dataset is

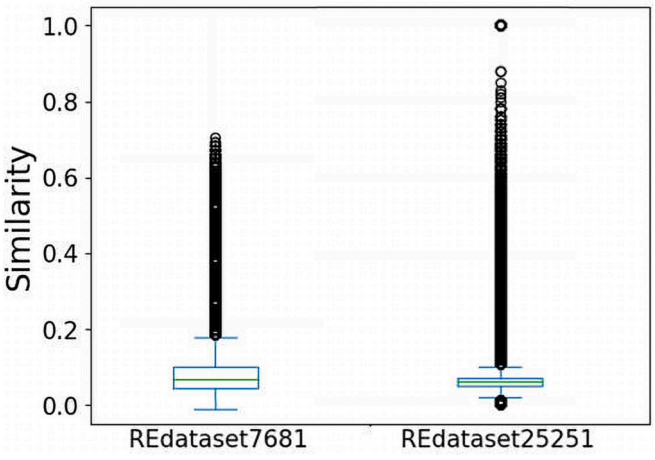

Fig. 3. The molecular similarity between the E_FMO and RE datasets. (REdataset7681: Tanimoto coefficient between Dataset75000 and REdataset7681, REdataset25251: Tanimoto coefficient between Dataset75000 and REdataset25251).

low, indicating that the vast majority of molecules in the dataset have minor similarity.

Table 1 presents detailed information on all datasets used in this study. To provide a more comprehensive understanding of the molecules within the dataset, we randomly selected 10 molecules from each dataset for visualization (see Fig. 4). We conducted stratified sampling

Table 1
Information of dataset used in this study.

Name	Size	Content
Dataset75000	75,000	E_FMO
REdataset7681	7681	RE
REdataset25251	25,251	RE

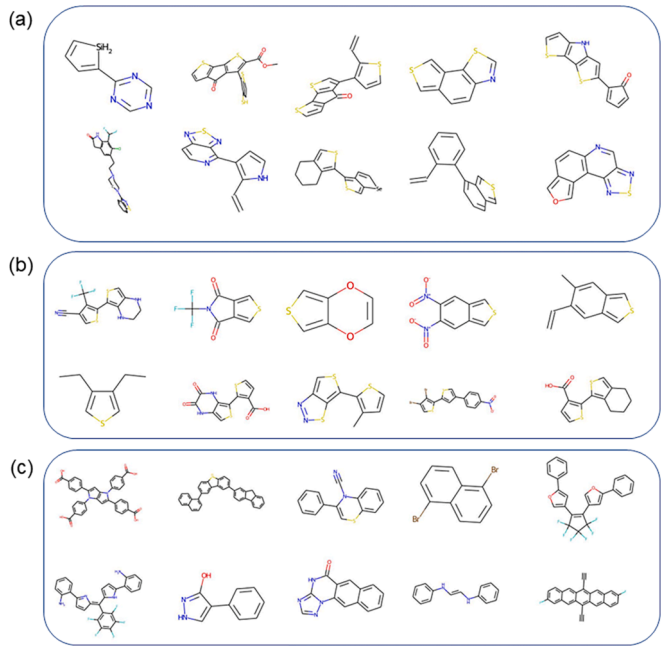

Fig. 4. Structural diagrams of 10 molecules from the datasets. (a) Dataset75000, (b) REdataset7681, (c) REdataset25251.

to extract small datasets, ranging in length from 100 to 7000, from the corresponding complete RE dataset, thus preserving the numerical distribution characteristics of the original dataset as closely as possible. We stratified the RE values at 0.1 intervals and proportionally selected sizes within each layer.

$$T = c/(a+b-c) \tag{2}$$

2.2. Computational methods and model selection

The network employs molecular fingerprints as the input. These are mathematical representation methods based on the molecule's atomic composition, bond connections, and other features, designed to describe its chemical structure. This method transforms chemical structures into numerical sequences, making them computer friendly. In this study, we utilized RDKit [81] to generate molecular fingerprints, which are one-dimensional arrays of specified lengths consisting of 0 or 1, from SDF files or SMILES (refer to Fig. 5). RDKit is a collection of cheminformatics and ML software that can calculate various types and lengths of molecular fingerprints, including 166 MACCS keys [82], extended circular fingerprints (ECFPs) [83], and Avalon fingerprints [84]. These molecular fingerprints serve to characterize the structural features of compounds. Previous studies [35,36] have used molecular fingerprints as inputs to predict a compound's energy, reinforcing the viability and effectiveness of using molecular fingerprints for energy calculations.

The multilayer perceptron (MLP), a feed-forward neural network representing the nonlinear mapping between an input vector and an output vector, can be optimized using the back-propagation algorithm. Its robust adaptive and self-learning function makes it well-suited for processing one-dimensional data. As a neural network, the MLP model shares similarities with the human brain's processing mechanism,

Fig. 5. Process of extracting molecular fingerprints.

functioning as an iterative and continuous abstraction process. The front layers of the network can be seen as feature extractors, and the extracted features are versatile [85]. In this study, the MLP model was interpreted as a nonlinear function $F(x)$ that takes molecular fingerprints as input x and produces E_FMO or RE as output y. During training, the function's coefficients (w_1, w_2, and w_3) were continuously adjusted to improve output performance. As shown in Eq. (3) and Eq. (4), f_1 and f_2 represent the activation functions used in the MLP model, introducing nonlinearity to the model. Compared to methods such as random forests, which divide molecular fingerprints into independent features, the MLP model is capable of extracting relationships between various local features of molecular fingerprints. Furthermore, the MLP network holds an advantage in fine-tuning the pre-training network using the target domain data when applying TL in the regression model. The purpose of this study is to leverage TL to address the problem of inaccurate predictions in the face of insufficient data. Therefore, the MLP model [86] was chosen for this study.

$$F(x) = w_3 * f_2(w_2 * f_1(w_1 * x)) \tag{3}$$

$$y = F(x) \tag{4}$$

The MLP model in this study comprises two hidden layers: one with 127 nodes and the other with 109 nodes. To accelerate the network training by reducing the internal covariate shift, a Batch Normalization layer was added between the input and the first hidden layer. The Continuously Differentiable Exponential Linear Units activation function, implemented using the EchoAI Python package [87], served as the activation function for the input and hidden layers. The output layer consists of a single node using a linear activation function. All middle layers had an L2 regularization coefficient set to 0.2 following optimization. All parameter settings mentioned above were based on the research published by Abarbanel [86]. To reduce training time, the EarlyStopping function in Keras was applied to halt training when no improvement was observed. The ReduceLROnPlateau Keras function was used to adjust the learning rate and minimize the number of training epochs. Finally, the model was implemented using Keras [88].

2.3. Model training

To evaluate the accuracy of the regression model's RE predictions, this study employed mean squared error (MSE) and root mean squared error (RMSE) using Eq. (5) and Eq. (6), respectively, where $\hat{y}_i$ represents the predicted value of the i^{th} observation, y_i represents the true value of the i^{th} observation, and n is the sample size. The R^2 score was used to measure the fit between the predicted values and true values. All experiments were performed on a computer with a 16 GB i7-11800H CPU. To ensure reliable results (including MSE, RMSE, R^2), a ten-fold cross validation was conducted to mitigate the influence of randomness. The small RE dataset was divided into 10 segments by the KFold function from the Scikit-learn Python package [89]. In each fold, nine parts were used for training and one for testing. The training set for each fold was trained using ML and TL, with results saved for two separate groups. The mean values of 10 iterations were utilized to assess the advantages of TL. The Wilcoxon signed-rank test [90] was applied to analyze the significant differences between paired ML and TL results from the same dataset. A *P-value* less than 0.05 indicated a significant difference between the two sets of results.

$$MSE = \sum (\hat{y}_i - y_i)^2 \Big/ n \tag{5}$$

$$RMSE = \sqrt{\sum (\hat{y}_i - y_i)^2 \Big/ n} \tag{6}$$

TL, known to enhance the prediction performance of the target domain using knowledge learned from the source domain and the corresponding task, was utilized in this study. E_FMO was the source

domain and RE was the target domain in the TL. The source domain, Dataset75000, was split evenly into a training set of 60,000 molecules and a test set of 15,000 molecules. In Algorithm 1, the pre-training process used 60,000 ECFP4 (2048) fingerprints, generated from the SDF file, and E_FMO as the input for the MLP model (Fig. 6. a). The output after a series of linear and nonlinear transformations was E_FMO. This pre-trained network was saved as a .h5 file. In Algorithm 2, this saved model was then reused to predict RE after fine-tuning (Fig. 6. b). A small portion of RE data was split into training and testing sets. The RE test set was predicted using the fine-tuned model trained with the RE training set and fingerprints. The fine-tuning process involved the input of the molecular fingerprint and RE into the pre-trained model, freezing the first layer, indicating that the input layer of the model was untrained, and fine-tuning the subsequent layers to output RE. This fine-tuned network was saved and reused to predict the RE. The prediction process involved feeding the molecular fingerprint of the corresponding compound into the network to predict the RE value (Fig. 6.c). This TL approach is based on feature transfer. Pre-training with a large dataset allowed the model to learn to extract E_FMO from molecular fingerprints, thereby enabling the faster and more accurate extraction of extract RE from molecular fingerprints using a smaller quantity of compounds.

Algorithm 1 Pre-training in TL

Input Source domain dataset containing E_FMO and molecular fingerprints. The dataset is stratified into training and testing datasets based on the values of E_FMO.
1. **do**
2. Adjust the parameters of the MLP model by training it with the training dataset.
3. Record the MSE on the validation dataset for the model after every epoch.
4. **While (EarlyStopping!= True)**
5. Select the best model with the lowest validation MSE and save the model as a. h5 file.
6. Use the best model to generate predictions on the testing dataset.

Output E_FMO

Algorithm 2 Fine-tuning in TL

Input Target domain dataset consisting RE and molecular fingerprints. The dataset is stratified into training and testing datasets based on the values of RE.
1. Load the saved model after pre-training.
2. Freeze the first layer of the model.
3. **do**

(*continued on next column*)

(*continued*)

Algorithm 2 Fine-tuning in TL

4. Fine-tune the unfrozen layers of the model by training it on the training dataset.
5. Record the MSE on the validation dataset for the model after every epoch.
6. **While (EarlyStopping!= True)**
7. Select the best model with the lowest validation MSE and save the model.
8. Use the best model to generate predictions on the testing dataset.

Output RE

3. Results and discussion

3.1. LUMO shows a better effect than HOMO and HOMO-LUMO gap as the source domain

The initial experiment sought to compare the effectiveness of TL when the source domain was either the HOMO or LUMO. Additionally, the HOMO-LUMO gap, calculated by subtracting the energy of the LUMO from the HOMO energy, was also used independently as the source domain for further comparison. The target domain, ranging from 100 to 4000, was uniformly selected from REdataset7681 via stratified sampling whenever the sample size was increased by 100. Due to space constraints, only the average R^2 value of ML and TL and the *P-value* of the Wilcoxon signed-rank test in three different source domains are shown in Table 2, with the specific test results presented in Appendix A. As illustrated in Table 2, the difference in R^2 between TL and ML was significant when the RE quantity was small. The interval in which the data volume of the small RE dataset was significant ranged from 300 to 3000 when the source domain was the LUMO, from 800 to 3000 when it was the HOMO, and again from 800 to 3000 when it was the HOMO-LUMO gap. Therefore, a network pre-trained on either the HOMO, LUMO, or HOMO-LUMO gap can enhance the network's prediction accuracy when the number of RE is limited. Thickened *P-values* denote a significant difference between two sets of results, while boldfaced R^2 represents a better result for the current data volume. Although the optimal R^2 after TL was achieved when the source domain was the HOMO-LUMO gap, it came second when the source domain was the LUMO. The difference in the R^2 values obtained using the LUMO and HOMO-LUMO gaps as the source domains was minimal. Moreover, when the quantity of RE data was small (e.g., between 300 and 700), there was no significant difference between TL and ML, with the HOMO-

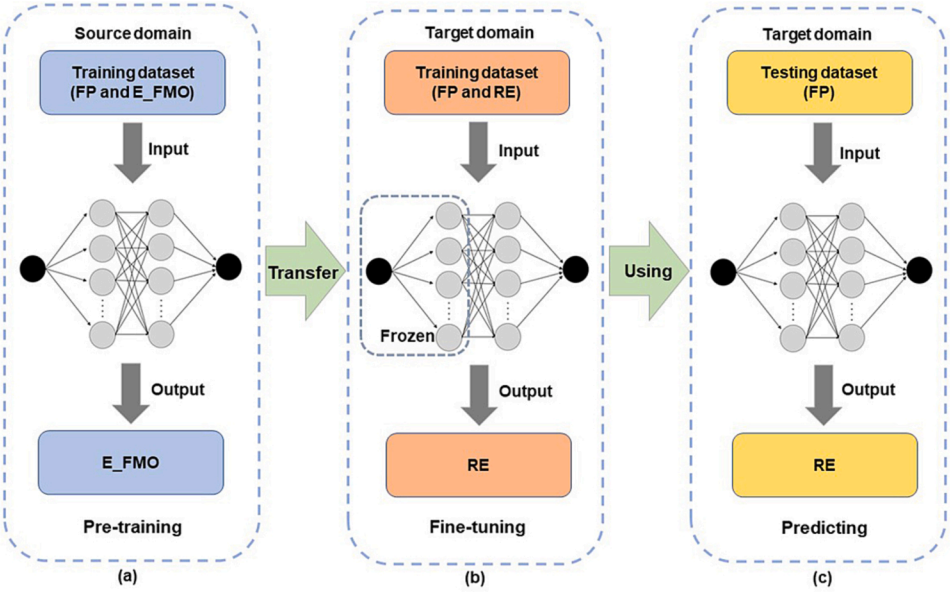

Fig. 6. Schematic diagram of proposed TL strategy. (FP: molecular fingerprint).

Table 2
Impact of different source domains on the performance of TL.

Data set	HOMO			LUMO			HOMO-LUMO gap		
	ML	TL	*P-value*[a]	ML	TL	*P-value*[a]	ML	TL	*P-value*[a]
100	−0.263335	**−0.029597**	0.130859	−0.291867	−0.255838	1.000000	−0.314673	−0.161548	0.431641
200	0.188186	**0.223014**	0.492188	0.144460	0.175997	0.193359	0.168661	0.129197	0.769531
300	0.252079	0.262895	0.556641	0.146856	0.276039	**0.009766**	0.210621	**0.279863**	0.083984
400	0.347110	**0.400498**	**0.037109**	0.293932	0.382255	**0.009766**	0.375593	0.372441	0.695313
500	0.339047	**0.410969**	0.064453	0.242531	0.390868	**0.019531**	0.386074	**0.399933**	0.322266
600	0.404809	**0.429557**	0.322266	0.355816	0.426846	**0.027344**	0.389338	0.424253	0.232422
700	0.370845	0.421940	0.130859	0.328088	**0.428346**	**0.005859**	0.335008	0.412022	0.064453
800	0.412094	0.459933	**0.001953**	0.374287	0.459103	**0.027344**	0.395207	**0.464293**	**0.009766**
900	0.377429	0.444682	**0.001953**	0.341296	0.451394	**0.001953**	0.414194	**0.455309**	0.193359
1000	0.414137	0.489961	**0.003906**	0.417685	**0.493334**	**0.009766**	0.414312	0.492658	**0.001953**
1100	0.419553	0.479476	**0.037109**	0.393460	0.483640	**0.003906**	0.422130	**0.486806**	**0.009766**
1200	0.393531	0.495721	**0.003906**	0.424683	**0.506077**	**0.005859**	0.417469	0.501262	**0.003906**
1300	0.463953	0.510797	**0.027344**	0.459224	0.513640	**0.005859**	0.460775	**0.519269**	**0.003906**
1400	0.481578	0.518000	0.083984	0.472947	0.520394	**0.003906**	0.460830	**0.525433**	**0.001953**
1500	0.456433	0.506532	0.009766	0.452574	0.503926	**0.001953**	0.467246	**0.512865**	**0.001953**
1600	0.464166	0.514810	**0.005859**	0.466469	**0.518004**	**0.001953**	0.477969	0.517811	**0.019531**
1700	0.495666	0.527131	**0.009766**	0.470460	**0.531142**	**0.005859**	0.483715	0.529701	**0.001953**
1800	0.495429	0.531925	**0.027344**	0.498410	0.537984	**0.001953**	0.499228	**0.544552**	**0.019531**
1900	0.495751	0.526953	**0.001953**	0.493217	0.531251	**0.001953**	0.497192	**0.533456**	**0.001953**
2000	0.506345	0.541447	**0.013672**	0.505343	0.536845	**0.001953**	0.516048	**0.545868**	**0.019531**
2100	0.500487	0.539531	**0.001953**	0.504885	0.540199	**0.013672**	0.500339	**0.541781**	**0.001953**
2200	0.531474	0.559184	**0.013672**	0.525608	0.561205	**0.001953**	0.539974	**0.564533**	**0.003906**
2300	0.522288	0.558149	**0.019531**	0.513004	0.560822	**0.009766**	0.521031	**0.562433**	**0.005859**
2400	0.532460	0.552289	0.064453	0.515458	0.554477	**0.013672**	0.528429	**0.557550**	**0.001953**
2500	0.524105	0.546357	**0.027344**	0.515881	0.551070	**0.027344**	0.523097	**0.554092**	**0.001953**
2600	0.524796	0.557297	**0.003906**	0.528463	0.556255	**0.037109**	0.530365	**0.562507**	**0.005859**
2700	0.553930	0.568416	0.105469	0.558727	0.570644	0.322266	0.560635	**0.573083**	0.193359
2800	0.537713	0.558304	**0.013672**	0.537426	0.560871	**0.037109**	0.545985	**0.564197**	**0.013672**
2900	0.536111	0.563059	**0.001953**	0.540170	0.563208	**0.027344**	0.541755	**0.565752**	**0.027344**
3000	0.549643	0.565294	**0.027344**	0.543261	0.568367	**0.013672**	0.555014	**0.568522**	0.105469
3100	0.568361	0.577258	0.193359	0.575761	0.580927	0.625000	0.571813	**0.581365**	0.130859
3200	0.576870	0.574686	0.921875	0.568729	0.579817	0.160156	0.565054	**0.582207**	**0.048828**
3300	0.562466	0.571894	0.130859	0.570036	0.579080	0.232422	0.565688	**0.579477**	**0.037109**
3400	0.562366	0.580947	**0.037109**	0.575235	0.582288	0.375000	0.573467	**0.587392**	0.064453
3500	0.565278	0.581911	**0.048828**	0.567009	0.585755	**0.019531**	0.567170	**0.585870**	**0.003906**
3600	0.575268	0.580703	0.492188	0.568362	0.579695	0.130859	0.570785	**0.581102**	**0.037109**
3700	0.585565	0.584686	1.000000	0.581048	0.587023	0.322266	0.578348	**0.591091**	**0.019531**
3800	0.588552	0.592167	0.625000	0.591842	0.597797	0.695313	0.589098	**0.599166**	0.105469
3900	0.576844	0.586367	0.193359	0.577330	**0.590417**	**0.027344**	0.579422	0.588693	0.064453
4000	0.572095	0.577202	0.492188	0.581613	**0.587442**	0.130859	0.572696	0.580735	0.232422

([a] *P-value* of Wilcoxon signed-rank test for R^2.).

LUMO gap as the source domain. Therefore, to better highlight the advantages of TL when data is scarce, the LUMO was chosen as the source domain in subsequent experiments after comprehensive consideration. In cases where RE is limited, using a network pre-trained with E_FMO and fine-tuning it through TL can enhance the accuracy of RE predictions. Under the current model, selecting the LUMO as the source domain for the TL yielded marginally better results. This can be explained as follows: in Marcus electron transfer theory, RE is related to the total energy when an ion transitions from an excited state back to the ground state, which is the sum of the two relaxation energies calculated from the excited and ground state. Although the HOMO is generally considered to have greater physical significance than the LUMO from the perspective of electron transfer and excitation, there is a certain linear relationship between the energy difference of the compounds and the LUMO. Therefore, a connection between the LUMO and RE may exist, and this relationship will be explored further in further studies.

3.2. TL significantly reduces the required size of the training dataset and computation time

Fig. 7 provides a clear comparison between the MLP model pre-trained on LUMO and the direct use of the MLP model for the prediction of RE. Regardless of whether TL or ML is applied, MSE, RMSE, and R^2 tend toward optimal results as the data size increases. However, the application of TL enables faster attainment of results that are close to optimal. When the dataset size was between 300 and 2500 using ML, the MSE ranged from 0.0193 to 0.0288 eV, RMSE ranged from 0.1381 to 0.1681 eV, and R^2 ranged from 0.2425 to 0.5256. After applying TL, the MSE further reduced to a range of 0.0178–0.0235 eV, the RMSE was reduced to a range of 0.1327–0.1512 eV, and R^2 increased to a range of 0.3908–0.5612 for the same data size. As depicted in Fig. 7.d, predicting the RE using ML significantly reduces prediction time. Training the network using 4000 RE took less than 30 s, and the time for predicting the RE of a compound using the trained network was negligible. Therefore, compared to traditional methods for calculating RE, ML offers considerable advantages in terms of computational time. Moreover, TL can significantly decrease the training time of the network as data size increases. The time spent by for TL does not include pre-training, as it is conducted separately from specific applications and belongs to the data pre-processing stage, and it does not contribute to the time cost of specific applications. During fine-tuning, the pre-trained network is directly loaded for training, and the network loading time can be disregarded. Because the EarlyStopping function is used to terminate network training when the prediction error no longer decreases, there may be some fluctuations in the training time curve. However, when the number of RE datasets was more than 2500, the gap between TL and ML narrowed, which is consistent with the principles of TL. This is because if the size of the dataset is close to the minimum amount of data required for a good prediction, the advantages of TL will be diminished. Nevertheless, this does not mean the effect of TL is unimpressive, as TL can still reduce the network training time. Differences in compound structures between different databases also impact the performance of TL.

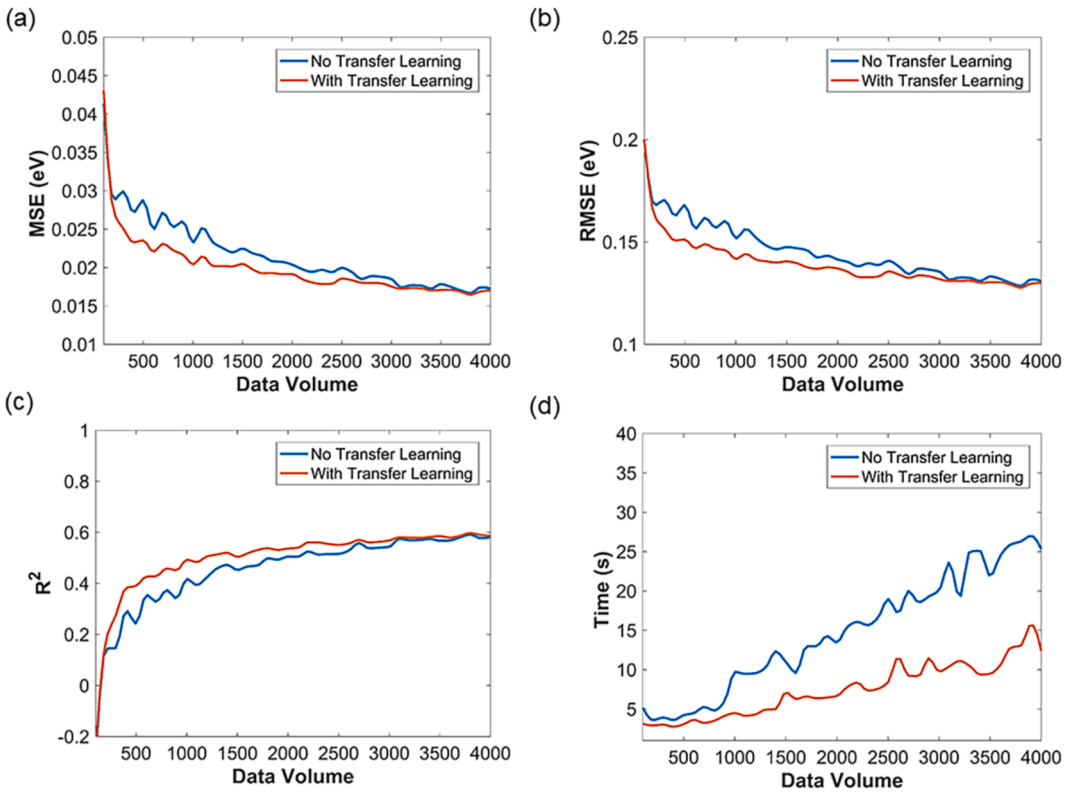

Fig. 7. Comparison of network performance between ML and TL when the source domain is LUMO. (a) Average MAE, (b) Average RMAE, (c) Average R^2, and (d) Average CPU run time of ten-fold cross validation between using ML and TL in REdataset7681. In (a), (b), (c), and (d), 'Data Volume' is the size of the small RE dataset.

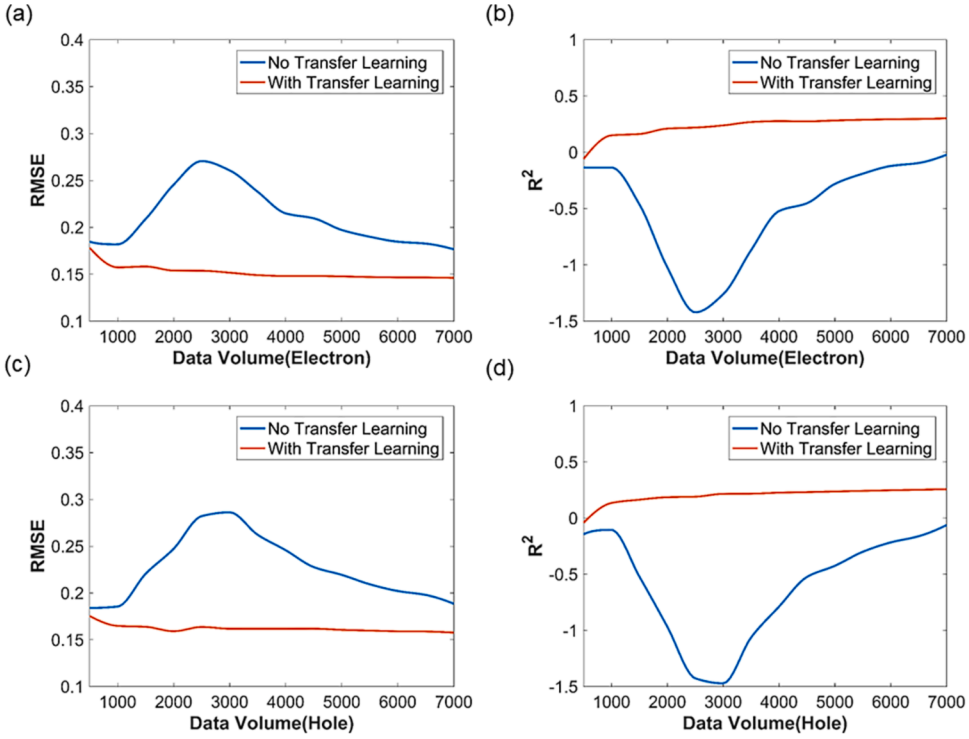

Fig. 8. Comparison of TL performance between target domains of electron and hole RE. (a) Average RMAE and (b) Average R^2 of ten-fold cross validation between using ML and TL in electron RE, (c) Average RMAE and (d) Average R^2 of ten-fold cross validation between using ML and TL in hole RE. In (a), (b), (c), and (d) 'Data Volume' is the size of the RE dataset.

However, it is assured that using the network pre-trained by E_FMO to predict the RE is efficient.

As previously mentioned, the REdataset7681 exclusively represents hole RE data. This study extends its investigation by focusing on the REdataset25251, which includes electron and hole RE. The aim is to explore how the differences between electron and hole RE affect the effectiveness of TL. For this, the LUMO is retained as the source domain, while the electron and hole RE from the same molecule are used as target domains. These datasets range from 500 to 7000 in size, and were selected from REdataset25251. Fig. 8 show that network performance is significantly improved after implementing TL, regardless of the target domain electron or hole RE. The RMSE is reduced to 0.1462– 0.1782 eV with an average improvement of 26% when the target domain is electron RE and to 0.1577– 0.1754 eV with an average improvement of 27% when the target domain is hole RE. Furthermore, R^2 increased from negative to positive for electron and hole RE. Therefore, this suggests a correlation between the LUMO and either the hole or electron RE based on the chemical structure. In the context of electronic devices such as OLEDs, electron and hole RE represent the energy required for an anion or cation to transfer from one molecule to another. Since electron and hole RE, as along with the LUMO, relate to charge transfer and chemical structure, applying TL from the LUMO to the electron and hole RE is plausible.

Despite the differences in calculation methods used in the source domain (B3LYP method with the 6-31G* basis set) and the target domain (LC-ωHPBE functionals and the Def2SVP basis set), TL demonstrates effectiveness, as seen in Fig. 8. This indicates that variation in quantum computational methods, including functionals and basis sets, have a negligible impact on the TL between the LUMO and the RE. It is also observed that network performance under ML worsens when the data increase from 500 to 2500. This could be due to the double descent phenomenon in ML, where the network performance first improves, then deteriorates, and finally improves again as the size of the dataset and training time increase. As demonstrated in Fig. 8, TL effectively mitigates this phenomenon. Although Fig. 7 indicates no significant difference in predicted results between the TL and ML for data quantities greater than 3500, Fig. 8 shows TL still maintains advantages when the data size reaches 7000. This suggests different RE datasets may impact performance. Significant differences in the data distributions across diverse RE datasets may necessitate more data to mitigate outliers' influence on the network and achieve effective predictions. However, TL significantly reduces the amount of data required for training, which is convenient for RE research and accelerates studies on the optical properties of materials.

Optoelectronic materials generally exhibit low RE values [19–21]. To illustrate the proposed method's effectiveness in predicting low RE, we conducted further experiments on electron and hole RE values ranging from 0 to 0.25 eV, extracted from the REdataset25251. The number of electron and hole RE within this range were 1478 and 4058, respectively. As shown in Fig. 9, ML and TL exhibit median prediction errors lower than 0.05 eV in the low RE dataset. Moreover, TL enhances prediction accuracy, demonstrated by a reduction in the maximum and an improvement in the average distribution of prediction errors. This effect is particularly significant with low hole RE. These findings indicate that TL offers notable advantages in predicting low RE, which could expedite the screening of materials with excellent optoelectronic properties. The optimization benefit of TL is even more evident in the prediction of low hole RE. This is attributed to the "double descent" phenomenon previously discussed concerning to ML in the context of REdataset25251. As observed in Fig. 8, ML may deliver poorer prediction performance as the dataset size increases from 1000 to 7000. Given that the number of hole RE surpasses the number of electron RE, ML's performance in predicting low hole RE tends to be subpar. However, TL effectively addresses this issue.

To better understand how TL can significantly reduce the amount of data required to achieve a comparable level of prediction accuracy, we conducted additional experiments. We used the Wilcoxon rank-sum test [91], a method for analyzing significant difference between unpaired data, to evaluate the MSE, RMSE, and R^2 values when using TL with a small dataset versus using ML with the total electron RE in REdataset25251. Table 3 displays the averages of the R^2, RMSE, and *P-value* results from the Wilcoxon rank-sum test, while Appendix A provides the detailed test results. As displayed in Table 3, the average RMSE and R^2 values resulting from ML's ten-fold cross validation on the complete REdataset25251 were 0.1462 eV and 0.3023, respectively. TL achieved a performance similar to ML on the entire REdataset25251 when the quantity of the small RE dataset exceeded 3500. Moreover, when using TL with 7000 RE, the RMSE was 0.1462 eV, and R^2 was 0.3013. This suggests that an equal effect, obtained in the total REdataset25251 using ML can be achieved using TL using only 28% of the complete dataset. Therefore, TL can significantly reduce the size of the training dataset required, thereby facilitating research in RE.

3.1.1. *Molecular fingerprinting has no significant impact on this study*

Lastly, we repeated the experiments using six different molecular fingerprints, with the LUMO in Dataset75000 as the source domain and the electron RE in REdataset25251 as the target domain. The goal was to investigate whether different molecular fingerprints would affect TL in this study. The average R^2 values for ML and TL, as well as the *P-values* of the Wilcoxon signed-rank test for different molecular fingerprints are presented in this Table 4. The specific test results can be found in Appendix A. Table 4 demonstrates that significant differences exist in the RMSE and R^2 values between TL and ML when different molecular fingerprints are used in the experiments. In the Wilcoxon signed-rank test, the *P-values* for the RMSE and R^2 are identical when the length of the molecular fingerprint exceeded 1024. This could be attributed to the fact that a molecular fingerprint longer than 1024 provides sufficient information for the computation of the LUMO and RE. Therefore, the

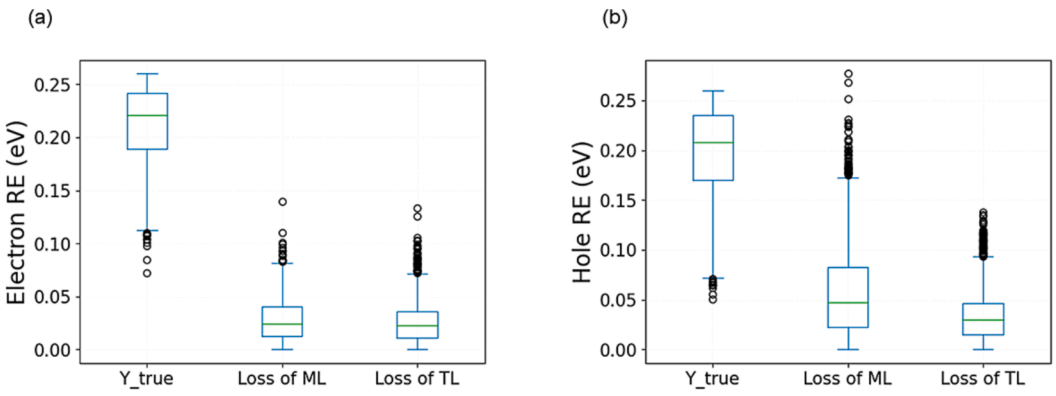

Fig. 9. Performance of ML and TL in predicting low RE. (Y_true: the true value of RE).

Table 3
Comparison of prediction performance between a small amount of data using TL and a complete dataset using ML.

Data set	Total[a]	RMSE		Total[a]	R^2	
		TL[b]	*P-value*[c]		TL[b]	*P-value*[c]
500		0.178231	**0.000507**		−0.059403	**0.000157**
1000		0.157322	**0.010165**		0.150939	**0.000881**
1500		0.158317	**0.004072**		0.162201	**0.000157**
2000		0.153964	**0.000881**		0.210513	**0.000285**
2500		0.153794	**0.001940**		0.219104	**0.000157**
3000	0.146192	0.151732	**0.008151**		0.238778	**0.002497**
3500		0.149139	0.150927	0.302339	0.268679	0.096304
4000		0.148167	0.173617		0.277067	0.256839
4500		0.148220	0.226476		0.274625	0.130570
5000		0.147728	0.150927		0.282720	0.049366
5500		0.147072	0.545350		0.289098	0.226476
6000		0.146689	0.820596		0.293784	0.289918
6500		0.146605	0.705457		0.296083	0.364346
7000		0.146169	0.939743		0.301278	0.939743

([a] The whole REdataset25251 using ML. [b] A small amount of RE using TL. [c] The *P-value* of Wilcoxon rank-sum test. Thickened *P-values* indicate a significant difference between the two sets of results.).

Table 4
Impact of diverse molecular fingerprints on the performance of TL.

	RMSE			R^2		
Fingerprint	**ML**	**TL**	***P-value***[b]	**ML**	**TL**	***P-value***[b]
MACCS (167)[a]	0.022850	0.022980	**0.003906**	0.254343	0.250054	**0.003906**
Avalon (512)[a]	0.020388	0.020318	**0.048828**	0.334602	0.336890	**0.048828**
ECFP2 (1024)[a]	0.020276	0.019949	**0.001953**	0.338464	0.349100	**0.001953**
ECFP2 (2048)[a]	0.020161	0.019057	**0.001953**	0.342169	0.378185	**0.001953**
ECFP3 (2048)[a]	0.020859	0.019756	**0.001953**	0.319375	0.355403	**0.001953**
ECFP4 (2048)[a]	0.021380	0.020215	**0.001953**	0.302339	0.340369	**0.001953**

([a] Length of the molecular fingerprint. [b] The *P-value* of Wilcoxon signed-rank test. Thickened *P-values* indicate a significant difference between two sets of results.).

model's performance is unaffected when the length of the molecular fingerprint used in the experiment surpasses 1024, thus indicating its insignificance.

4. Conclusions

This study addressed the issue of subpar network performance in the face of insufficient RE data through the use of TL. The ensuing experiments yielded the following conclusions: 1) Utilizing TL with E_FMO as the source domain can enhance the accuracy of network prediction when there is a scarcity of training data. This approach significantly reduces the size of the required training dataset, thereby facilitating RE research and expediting the study of material optical properties; 2) The TL strategy applied in this study proved effective for both the electron and hole RE within the target domain; 3) TL offers certain advantages in predicting low RE; and 4) The impact of molecular fingerprint length on TL was negligible in this experiment once its length exceeded 1024. Notably, our network merely requires molecular fingerprints as inputs to predict RE. The network proposed here simplifies the input process, thereby making RE prediction via ML more convenient. Although the Tanimoto coefficient indicates minor molecular similarity between the E_FMO and RE datasets, TL is still effective. It demonstrates that molecular fingerprints contain some information shared by RE and E_FMO. TL can learn how to extract this information and reuse it for predicting RE. Moreover, it is also effective to predict the RE using a model pretrained with the HOMO, LUMO, or HOMO-LUMO gap, as demonstrated in Table 2. This method is significant for predicting other chemical quantities related to compound structure. Future research endeavors will consider the inclusion of other compound feature information, such as compound surface maps, to further enhance prediction accuracy.

5. Data availability.

All datasets mentioned in this paper can be found, in the online version, at https://data.mendeley.com/datasets/xg96w33syr/1, https://doi.org/10.17632/xg96w33syr.1.

CRediT authorship contribution statement

Xushi Zhang: Data curation, Writing – original draft, Writing – review & editing, Software. **Guodong Ye:** Conceptualization, Data curation, Writing – original draft. **Chuanxue Wen:** Conceptualization, Methodology, Writing – original draft. **Zhisheng Bi:** Conceptualization, Methodology, Writing – original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work is supported by the Guangzhou Education Science Planning Project (No. 202214340 and No. 2023TLKC010),the Discipline Construction Project of Guangzhou Medical University (No. 02-445-2301244XM and No. 01-408-2201015) and the Industry-university Cooperative Education Project (No. 201902120032).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.commatsci.2023.112361.

References

[1] S.R. Forrest, The path to ubiquitous and low-cost organic electronic appliances on plastic, Nature. 428 (2004) 911–918, https://doi.org/10.1038/nature02498.

[2] M.E. Gershenson, V. Podzorov, A.F. Morpurgo, Colloquium: Electronic transport in single-crystal organic transistors, Rev. Mod. Phys. 78 (2006) 973–989, https://doi.org/10.1103/RevModPhys.78.973.

[3] S. Sergeyev, W. Pisula, Y.H. Geerts, Discotic liquid crystals: A new generation of organic semiconductors, Chem. Soc. Rev. 36 (2007) 1902–1929, https://doi.org/10.1039/b417320c.

[4] J.D. Myers, J. Xue, Organic semiconductors and their applications in photovoltaic devices, Polym. Rev. 52 (2012) 1–37, https://doi.org/10.1080/15583724.2011.644368.

[5] S.C. Lo, P.L. Burn, Development of dendrimers: Macromolecules for use in organic light-emitting diodes and solar cells, Chem. Rev. 107 (2007) 1097–1116, https://doi.org/10.1021/cr050136l.

[6] Y. Kim, S.A. Choulis, J. Nelson, et al., Composition and annealing effects in polythiophene/fullerene solar cells, J. Mater. Sci. 40 (2005) 1371–1376, https://doi.org/10.1007/s10853-005-0568-0.

[7] M. Zhang, X. Guo, W. Ma, et al., A polythiophene derivative with superior properties for practical application in polymer solar cells, Adv. Mater. 26 (2014) 5880–5885, https://doi.org/10.1002/adma.201401494.

[8] Z.G. Zhang, S. Zhang, J. Min, et al., Side chain engineering of polythiophene derivatives with a thienylene–vinylene conjugated side chain for application in polymer solar cells, Macromolecules. 45 (2012) 2312–2320, https://doi.org/10.1021/ma2026463.

[9] Z. Bao, A.J. Lovinger, Soluble regioregular polythiophene derivatives as semiconducting materials for field-effect transistors, Chem. Mater. 11 (1999) 2607–2612, https://doi.org/10.1021/cm990290m.

[10] R. Porrazzo, S. Bellani, A. Luzio, et al., Field-effect and capacitive properties of water-gated transistors based on polythiophene derivatives, APL Mater. 3 (2015), 014905, https://doi.org/10.1063/1.4900888.

[11] F. Wang, H. Gu, T.M. Swager, Carbon nanotube/polythiophene chemiresistive sensors for chemical warfare agents, J. Am. Chem. Soc. 130 (2008) 5392–5393, https://doi.org/10.1021/ja710795k.

[12] P. Schottland, M. Bouguettaya, C. Chevrot, Soluble polythiophene derivatives for NO2 sensing applications, Synth. Met. 102 (1999) 1325, https://doi.org/10.1016/S0379-6779(98)01043-1.

[13] L. Wang, Q. Feng, X. Wang, et al., A novel polythiophene derivative as a sensitive colorimetric and fluorescent sensor for anionic surfactants in water, New J. Chem. 36 (2012) 1897–1901, https://doi.org/10.1039/c2nj40460e.

[14] B.H. Barboza, O.P. Gomes, A. Batagin-Neto, Polythiophene derivatives as chemical sensors: A dft study on the influence of side groups, J. Mol. Model. 27 (2021) 17, https://doi.org/10.1007/s00894-020-04632-w.

[15] S.K. Kang, J.H. Kim, J. An, et al., Synthesis of polythiophene derivatives and their application for electrochemical dna sensor, Polym. J. 36 (2004) 937–942, https://doi.org/10.1295/polymj.36.937.

[16] A.L. Ding, J. Pei, Y.H. Lai, et al., Phenylene-functionalized polythiophene derivatives for light-emitting diodes: their synthesis, characterization and properties, J. Mater. Chem. 11 (2001) 3082–3086, https://doi.org/10.1039/b103717j.

[17] S.M. Sze, Y. Li, K.K. Ng, Physics of semiconductor devices, John Wiley & Sons, 2021.

[18] D. Moia, V. Vaissier, I. López-Duarte, et al., The reorganization energy of intermolecular hole hopping between dyes anchored to surfaces, Chem. Sci. 5 (2014) 281–290, https://doi.org/10.1039/C3SC52359D.

[19] D.A. da Silva Filho, E.G. Kim, J.L. Brédas, Transport properties in the rubrene crystal: Electronic coupling and vibrational reorganization Energy, Adv. Mater. 17 (2005) 1072–1076, https://doi.org/10.1002/adma.200401866.

[20] R. Saxena, V.R. Nikitenko, I.I. Fishchuk, et al., Role of the reorganization energy for charge transport in disordered organic semiconductors, Phys. Rev. B. 103 (2021), 165202, https://doi.org/10.1103/PhysRevB.103.165202.

[21] S. Fatayer, B. Schuler, W. Steurer, et al., Reorganization energy upon charging a single molecule on an insulator measured by atomic force microscopy, Nat. Nanotechnol. 13 (2018) 376–380, https://doi.org/10.1038/s41565-018-0087-1.

[22] C.P. Hsu, Reorganization energies and spectral densities for electron transfer problems in charge transport materials, Phys. Chem. Chem. Phys. 22 (2020) 21630–21641, https://doi.org/10.1039/d0cp02994g.

[23] B. Zhang, Y. Xu, L. Zhu, et al., Theoretical evaluation of the influence of molecular packing mode on the intramolecular reorganization energy of oligothiophene molecules, Polymers. 10 (2017) 30, https://doi.org/10.3390/polym10010030.

[24] S. Atahan-Evrenk, A quantitative structure–property study of reorganization energy for known p-type organic semiconductors, RSC Adv. 8 (2018) 40330–40337, https://doi.org/10.1039/c8ra07866a.

[25] H. Imahori, H. Yamada, D.M. Guldi, et al., Comparison of reorganization energies for intra-and intermolecular electron transfer, Angew. Chem. 114 (2002) 2450–2453, https://doi.org/10.1002/1521-3757(20020703)114:13<2450::AID-ANGE2450>3.0.CO;2-R.

[26] D.P. McMahon, A. Troisi, Evaluation of the external reorganization energy of polyacenes, J. Phys. Chem. Lett. 1 (2010) 941–946, https://doi.org/10.1021/jz1001049.

[27] A.B. Myers, Resonance Raman intensities and charge-transfer reorganization energies, Chem. Rev. 96 (1996) 911–926, https://doi.org/10.1021/cr950249c.

[28] R.A. Marcus, On the theory of oxidation-reduction reactions involving electron transfer. I, J. Chem. Phys. 24 (1956) 966–978, https://doi.org/10.1063/1.1742723.

[29] R.A. Marcus, On the theory of oxidation-reduction reactions involving electron transfer. II. applications to data on the rates of isotopic exchange reactions, J. Chem. Phys. 26 (1957) 867–871, https://doi.org/10.1063/1.1743423.

[30] R.A. Marcus, On the theory of oxidation-reduction reactions involving electron transfer. III. applications to data on the rates of organic redox reactions, J. Chem. Phys. 26 (1957) 872–877, https://doi.org/10.1063/1.1743424.

[31] R.A. Marcus, On the theory of electron-transfer reactions. vi. unified treatment for homogeneous and electrode reactions, J. Chem. Phys. 43 (1965) 679–701, https://doi.org/10.1063/1.1696792.

[32] G.R. Hutchison, M.A. Ratner, T.J. Marks, Hopping transport in conductive heterocyclic oligomers: reorganization energies and substituent effects, J. Am. Chem. Soc. 127 (2005) 2339–2350, https://doi.org/10.1021/ja0461421.

[33] J. Cornil, D. Beljonne, J.-P. Calbert, et al., Interchain interactions in organic π-conjugated materials: Impact on electronic structure, optical response, and charge transport, Adv. Mater. 13 (2001) 1053–1067, https://doi.org/10.1002/1521-4095(200107)13:14<1053::AID-ADMA1053>3.0.CO;2-7.

[34] S.S. Zade, M. Bendikov, Study of hopping transport in long oligothiophenes and oligoselenophenes: Dependence of reorganization energy on chain length, Chemistry. 14 (2008) 6734–6741, https://doi.org/10.1002/chem.200701182.

[35] O.D. Abarbanel, G.R. Hutchison, Machine learning to accelerate screening for marcus reorganization energies, J. Chem. Phys. 155 (2021), 054106, https://doi.org/10.1063/5.0059682.

[36] F. Pereira, K. Xiao, D.A.R.S. Latino, et al., Machine learning methods to predict density functional theory B3LYP energies of HOMO and LUMO orbitals, J. Chem. Inf. Model. 57 (2017) 11–21, https://doi.org/10.1021/acs.jcim.6b00340.

[37] J. Behler, Neural network potential-energy surfaces in chemistry: A tool for large-scale simulations, Phys. Chem. Chem. Phys. 13 (2011) 17930–17955, https://doi.org/10.1039/c1cp21668f.

[38] J. Behler, Constructing high-dimensional neural network potentials: A tutorial review, Int. J. Quantum Chem. 115 (2015) 1032–1050, https://doi.org/10.1002/qua.24890.

[39] J. Behler, Perspective: Machine learning potentials for atomistic simulations, J. Chem. Phys. 145 (2016), 170901, https://doi.org/10.1063/1.4966192.

[40] P.O. Dral, Quantum chemistry in the age of machine learning, J. Phys. Chem. Lett. 11 (2020) 2336–2347, https://doi.org/10.1021/acs.jpclett.9b03664.

[41] O.A. von Lilienfeld, K.R. Müller, A. Tkatchenko, Exploring chemical compound space with quantum-based machine learning, Nat. Rev. Chem. 4 (2020) 347–358, https://doi.org/10.1038/s41570-020-0189-9.

[42] F. Noé, A. Tkatchenko, K.R. Müller, et al., Machine learning for molecular simulation, Annu. Rev. Phys. Chem. 71 (2020) 361–390, https://doi.org/10.1146/annurev-physchem-042018-052331.

[43] S. Atahan-Evrenk, F.B. Atalay, Prediction of intramolecular reorganization energy using machine learning, J. Phys. Chem. A. 123 (2019) 7855–7863, https://doi.org/10.1021/acs.jpca.9b02733.

[44] M. Misra, D. Andrienko, B. Baumeier, et al., Toward quantitative structure–property relationships for charge transfer rates of polycyclic aromatic hydrocarbons, J. Chem. Theory. Comput. 7 (2011) 2549–2555, https://doi.org/10.1021/ct200231z.

[45] H. Sahu, H. Ma, Unraveling correlations between molecular properties and device parameters of organic solar cells using machine learning, J. Phys. Chem. Lett. 10 (2019) 7277–7284, https://doi.org/10.1021/acs.jpclett.9b02772.

[46] M. Rinderle, W. Kaiser, A. Mattoni, et al., Machine-learned charge transfer integrals for multiscale simulations in organic thin films, J. Phys. Chem. C. 124 (2020) 17733–17743, https://doi.org/10.1021/acs.jpcc.0c04355.

[47] D. Padula, J.D. Simpson, A. Troisi, Combining electronic and structural features in machine learning models to predict organic solar cells properties, Mater. Horiz. 6 (2019) 343–349, https://doi.org/10.1039/C8MH01135D.

[48] D. Padula, A. Troisi, Concurrent optimization of organic donor–acceptor pairs through machine learning, Adv. Energy Mater. 9 (2019) 1902463, https://doi.org/10.1002/aenm.201902463.

[49] C. Chen, Y. Zuo, W. Ye, et al., A critical review of machine learning of energy materials, Adv. Energy Mater. 10 (2020) 1903242, https://doi.org/10.1002/aenm.201903242.

[50] T. Sato, T. Honma, S. Yokoyama, Combining machine learning and pharmacophore-based interaction fingerprint for in silico screening, J. Chem. Inf. Model. 50 (2010) 170–185, https://doi.org/10.1021/ci900382e.

[51] J. Vamathevan, D. Clark, P. Czodrowski, et al., Applications of machine learning in drug discovery and development, Nat. Rev. Drug Discov. 18 (2019) 463–477, https://doi.org/10.1038/s41573-019-0024-5.

[52] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (2010) 1345–1359, https://doi.org/10.1109/TKDE.2009.191.

[53] K. Weiss, T.M. Khoshgoftaar, D.D. Wang, A survey of transfer learning, J. Big Data. 3 (2016) 1–40, https://doi.org/10.1186/s40537-016-0043-6.

[54] N. Dandu, L. Ward, R.S. Assary, et al., Quantum-chemically informed machine learning: Prediction of energies of organic molecules with 10 to 14 non-hydrogen atoms, J. Phys. Chem. A. 124 (2020) 5804–5811, https://doi.org/10.1021/acs.jpca.0c01777.

[55] X. Li, D. Fourches, Inductive transfer learning for molecular activity prediction: Next-gen qsar models with MolPMoFiT, Journal of Cheminformatics 12 (1) (2020) 1–15, https://doi.org/10.1186/s13321-020-00430-x.
[56] R.S. Simões, V.G. Maltarollo, P.R. Oliveira, et al., Transfer and multi-task learning in qsar modeling: Advances and challenges, Front. Pharmacol. 9 (2018) 74, https://doi.org/10.3389/fphar.2018.00074.
[57] J.S. Smith, B.T. Nebgen, R. Zubatyuk, et al., Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning, Nat. Commun. 10 (2019) 2903, https://doi.org/10.1038/s41467-019-10827-4.
[58] E.O. Pyzer-Knapp, K. Li, et al., Learning from the harvard clean energy project: The use of neural networks to accelerate materials discovery, Adv. Funct. Mater. 25 (2015) 6495–6502, https://doi.org/10.1002/adfm.201501919.
[59] S.A. Siddiqui, A. Al-Hajry, M.S. Al-Assiri, Ab initio investigation of 2, 2′-bis (4-trifluoromethylphenyl)-5, 5′-bithiazole for the design of efficient organic field-effect transistors, Int. J. Quantum Chem. 116 (2016) 339–345, https://doi.org/10.1002/qua.25034.
[60] G. Serdaroglu, N. Uludag, Concise total synthesis of (±)-aspidospermidine and computational study: FT-IR, NMR, NBO, NLO, FMO, MEP diagrams, J. Mol. Struct. 1166 (2018) 286–303, https://doi.org/10.1016/j.molstruc.2018.04.050.
[61] W.C. Chen, Y.C. Cheng, Elucidating the magnitude of internal reorganization energy of molecular excited states from the perspective of transition density, J. Phys. Chem. A. 124 (2020) 7644–7657, https://doi.org/10.1021/acs.jpca.0c06482.
[62] B.C. Lin, C.P. Cheng, Z.P.M. Lao, Reorganization energies in the transports of holes and electrons in organic amines in organic electroluminescence studied by density functional theory, J. Phys. Chem. A. 107 (2003) 5241–5251, https://doi.org/10.1021/jp0304529.
[63] C. Lee, W. Yang, R.G. Parr, Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density, Phys. Rev. B Condens. Matter. 37 (1988) 785–789, https://doi.org/10.1103/physrevb.37.785.
[64] V.A. Rassolov, J.A. Pople, M.A. Ratner, et al., 6–31G* basis set for atoms K through Zn, J. Chem. Phys. 109 (1998) 1223–1229, https://doi.org/10.1063/1.476673.
[65] J. Aires-de-Sousa, A.R.S. Diogo Energies of the HOMO and LUMO Orbitals for 111725 Organic Molecules Calculated by DFT. Figs Hare. (2016) https://doi.org/10.6084/m9.figshare.3384184.v1, [accessed 15 October 2022].
[66] J. Chem., ChemAxon: Chem-bioinformatics software for the next generation of scientists. http://www.chemaxon.com (2015) 15.4.6, [accessed 23 October 2022].
[67] OpenBabel, Open Babel: The Open Source Chemistry Toolbox, 2016. http://openbabel.org,[accessed 13 October 2022].
[68] O.D. Abarbanel, G.R. Hutchison, The dataset of reorganization energy, 2021. https://github.com/hutchisonlab/ReorganizationEnergy/tree/main/data, [accessed 10 October 2022].
[69] T. Schmidt, S. Kümmel, The influence of one-electron self-interaction on d-electrons, Computation. 4 (2016) 33, https://doi.org/10.3390/computation4030033.
[70] C.J. Cramer, D.G. Truhlar, Density functional theory for transition metals and transition metal chemistry, Phys. Chem. Chem. Phys. 11 (2009) 10757–10816, https://doi.org/10.1039/b907148b.
[71] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, et al., The harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid, J. Phys. Chem. Lett. 2 (2011) 2241–2251, https://doi.org/10.1021/jz200866s.
[72] The dataset with chromophore descriptors for electronic, optical and redox properties computed with DFT, 2022, [Online; accessed 1-Dec-2022]. https://oscar.as.uky.edu/datasets.
[73] Q. Ai, V. Bhat, S.M. Ryno, et al., OCELOT: An infrastructure for data-driven research to discover and design crystalline organic semiconductors, J. Chem. Phys. 154 (2021), 174705, https://doi.org/10.1063/5.0048714.
[74] T.M. Henderson, A.F. Izmaylov, G. Scalmani, et al., Can short-range hybrids describe long-range-dependent properties? J. Chem. Phys. 131 (2009), 044108 https://doi.org/10.1063/1.3185673.
[75] F. Weigend, R. Ahlrichs, Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy, Phys. Chem. Chem. Phys. 7 (2005) 3297–3305, https://doi.org/10.1039/b508541a.
[76] R. Baer, E. Livshits, U. Salzner, Tuned range-separated hybrids in density functional theory, Annu. Rev. Phys. Chem. 61 (2010) 85–109, https://doi.org/10.1146/annurev.physchem.012809.103321.
[77] K. Burke, J. Werschnik, E.K.U. Gross, Time-dependent density functional theory: Past, present, and future, J. Chem. Phys. 123 (2005) 62206, https://doi.org/10.1063/1.1904586.
[78] M.E. Casida, Time-dependent density-functional theory for molecules and molecular solids, J. Mol. Struct. THEOCHEM. 914 (2009) 3–18, https://doi.org/10.1016/j.theochem.2009.08.018.
[79] D. Bajusz, A. Rácz, K. Héberger, Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform 7 (2015) 1–13, https://doi.org/10.1186/s13321-015-0069-3.
[80] T. Le, R. Winter, F. Noé, et al., Neuraldecipher–reverse-engineering extended-connectivity fingerprints (ECFPs) to their molecular structures, Chem. Sci. 11 (2020) 10378–10389, https://doi.org/10.1039/D0SC03115A.
[81] RDKit, Online, RDKit: Open-Source Cheminformatics, 2021. http://www.rdkit.org, [accessed 8 October 2022].
[82] J.L. Durant, B.A. Leland, D.R. Henry, et al., Reoptimization of mdl keys for use in drug discovery, J. Chem. Inf. Comput. Sci. 42 (2002) 1273–1280, https://doi.org/10.1021/ci010132r.
[83] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (2010) 742–754, https://doi.org/10.1021/ci100050t.
[84] P. Gedeck, B. Rohde, C. Bartels, QSAR–how good is it in practice? comparison of descriptor sets on an unbiased cross section of corporate data sets, J Chem Inf Model. 46 (2006) 1924–1936, https://doi.org/10.1021/ci050413p.
[85] C. Tan, F. Sun, T. Kong, et al., A survey on deep transfer learning, Artificial Neural Networks and Machine Learning–ICANN 2018 Proceedings, Part III 27, Springer International Publishing, 2018, pp. 270–279.
[86] O.D. Abarbanel, G.R. Hutchison, The code of the MLP model used in this paper, 2021. https://github.com/hutchisonlab/ReorganizationEnergy/tree/main/models, [accessed 10 October 2022].
[87] Echo, Echo: Python package containing all custom layers used in Neural Networks, 2020. https://github.com/digantamisra98/Echo, [accessed 13 October 2022].
[88] Keras, Keras: Deep Learning for Humans, 2015. https://keras.io,[accessed 13 October 2022].
[89] Scikit-Learn, Scikit-Learn: Machine Learning in Python, 2022. https://scikit-learn.org, [accessed 10 October 2022].
[90] D. Rey, Neuhäuser, M, Wilcoxon-Signed-Rank Test. in: M. Lovric, (Eds.) International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg. 2011. pp. 1658-1659. https://doi.org/10.1007/978-3-642-04898-2_616.
[91] W.Haynes, Wilcoxon Rank Sum Test. in: W. Dubitzky, O. Wolkenhauer, K.H. Cho, H. Yokota (Eds.) Encyclopedia of Systems Biology, Springer New York, New York, NY, 2013. pp. 2354-2355. https://doi.org/10.1007/978-1-4419-9863-7_1185.