PAPER

Check for updates

An artificial neural network using multi-head intermolecular attention for predicting chemical reactivity of organic materials†

Jaekyun Yoo,[a] Byunghoon Kim,[a] Byungju Lee,[b] Jun-hyuk Song[a] and Kisuk Kang [iD] *[ac]

Selecting functional materials that are chemically compatible with each other is a prerequisite for the assembly of multi-component systems and is crucial for their long-term system stability. In the design of new organic-based batteries, one of the promising post-lithium-ion battery systems, the exploration of organic compounds for the electrode and electrolyte should consider not only their intrinsic electrochemical activity/stability but also the compatibility among the constituting components. Herein, we report an extensive scheme of predicting the chemical reactivities of any combination of two organic compounds by employing the so-called Intermolecular Reaction Rate Network (ImRRNet). This new artificial neural network (ANN) platform exploits the novel intermolecular multi-head attention method to predict the precise reaction rate constant between two organic chemicals and was trained with a large chemical space of 175 987 datasets on nucleophilicity and electrophilicity. The intermolecular multi-head attention method successfully identified the local substructure that primarily determines the chemical reactivity of organic molecules by providing a greater attention score in the specific position. The prediction accuracy of ImRRNet was observed to be remarkably higher (mean absolute error of 0.5760) than that of other previous ANN models (>0.94), validating its efficacy for practical employment in the design of multi-component organic-based rechargeable batteries.

10th anniversary statement

We are excited to be contributing to the 10th-anniversary issue of the *Journal of Materials Chemistry A*, a prominent journal in the research field of energy storage materials. We remember publishing our first paper in the journal about ten years ago about the transition metal migration in Li-excess cathode materials and its interplay with the voltage fade. Since then, significant efforts have been dedicated to understanding the degradation mechanisms of these materials in order to design high-energy density and long-cycle-life electrodes. Thanks to the research efforts over the past decade, it is now understood that regulating transition metal migration is a key to developing high-performance Li-excess materials. Our current research in this issue focuses on an artificial neural network model that can efficiently predict reaction rate constants with minimal computational cost, thus accelerating the speed of the material discovery process. Our novel approach will provide cost-effective and rapid predictions on the reactivity of various organic species and is expected to expedite the development of next-generation organic components in batteries. We hope this work will have a meaningful impact on the energy material research community in the next ten years, helping further to advance the field of energy storage materials.

Introduction

The upsurge in the demand for electric vehicles prompted the massive production of lithium-ion batteries (LIBs) in recent years, which is now projected to grow even faster in the upcoming years.[1,2] This explosive growth casts questions on the sustainability of current battery manufacture and chemistry, which is based on transition-metal compounds. Such compounds have limited reserves, presenting a potential supply problem, making production costs unreliable/high and the large-scale use of these compounds unsustainable.[3,4] Therefore, post-LIB systems such as rechargeable organic batteries that exploit redox-active organic electrodes or a catholyte/anolyte have garnered tremendous attention as sustainable battery systems.[5–9] These redox-active organic materials are mostly composed of earth-abundant carbon, oxygen, hydrogen, and nitrogen, with remarkable recent progress in the electrode performance, which can sometimes rival that of their transition-metal-containing counterparts according to the latest results.[10,11] More recently, extensive investigations have introduced various new high-performance redox-active organic

[a]*Department of Materials Science and Engineering, Research Institute of Advanced Materials (RIAM), Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Republic of Korea. E-mail: matlgen1@snu.ac.kr*

[b]*Center for Computational Science, Korea Institute of Science and Technology (KIST), Hwarangno 14-gil 5, Seongbuk-gu, Seoul 136-791, Korea*

[c]*Center for Nanoparticle Research at Institute for Basic Science (IBS), 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d2ta07660h

materials, which continue to open up the pathway toward sustainable batteries. However, implementing these redox-active organic electrodes in practical battery systems requires a system-level approach that considers the component compatibility of the organic cathode, anode, and electrolyte to reach desirable battery performance.[12–14] One of the major bottlenecks is that the organic species participating in redox reactions are highly susceptible to side reactions with other components in the batteries, leading to severe deterioration of energy retention.[5,15] In particular, organic redox-flow batteries, one of the major battery systems that employ organic compounds,[16,17] rely on the dynamic transport of positive and negative redox-active molecules at various charged states in supporting organic electrolytes, which induces complex physical and chemical interactions. Any chemical incompatibility among these molecules would lead to the gradual loss of the redox activity of the active organic component, which will be exacerbated along with its intrinsic electrochemical instability.[10,15]

Predicting chemical reactivities and extracting a general rule that governs them have been the central topics of the chemistry field from the past,[18,19] providing knowledge that is beneficial in the material design of organic redox materials for battery applications. Traditional ways to evaluate chemical reactivity have mainly employed transition-state theory[20,21] or the regression method using reaction indices such as hardness/softness, vertical ionization energy, and electron affinity.[22,23] However, despite reasonable accuracies, these traditional approaches are extremely laborious either experimentally[24] or computationally[25] and sometimes require substantial prior chemical knowledge on the subject materials. In addition, reaction-indice-based regression methods have a clear limitation in that only rough predictions are allowed for a relatively narrow range of chemistry.[22,26] In this respect, the recent development of machine learning (ML) techniques is promising, making fast and comprehensive predictions of organic reactivities possible through the exploration of large chemical databases. According to recent studies, the application of ML models could allow highly accurate prediction of the reaction rate constant ($R^2 > 0.8$)[27,28] and chemical selectivity (accuracy > 85%).[29]

Although ML-based predictions are expected to be efficient with higher precision at lower cost compared with conventional approaches, earlier studies were limited to only a relatively narrow chemical space with a handful of training datasets.[30,31] In addition, previous high-precision models often required additional quantum chemical computations for training,[27,32] which are typically costly in terms of computational resources. More importantly, the reactivity prediction using conventional ML methods is performed in a black-box manner, which hampers rational elucidation of the chemical origins of the reactivity between reactants. These aspects have impeded the practical employment of ML-based material screening/design and predictions of their chemical compatibility in organic-battery systems. Herein, we address these issues by implementing multi-head attention methods in the artificial neural network (ANN) model, which simultaneously (i) predicts reaction rate constants and (ii) elucidates relevant molecular substructures of significance to the reactivities, without any quantum chemical calculation. Our model, the so-called Intermolecular Reaction Rate Network (ImRRNet), exploits multi-head attention between reactant molecules, which takes account of the weight importance of molecular substructures where the major reactivity appears during the reaction. We demonstrate how ImRRNet can offer chemical intuitions regarding reactivity by visualizing the attention score[33] between molecular substructures. Comparison of performances reveals that ImRRNet presents a far better prediction capability with lower MAEs (∼0.5760) than those achieved using common ANN models (>0.94), which translates into more than twice the accuracy in the estimation of the reaction rate constant. We believe that the architecture of ImRRNet proposed here can be effectively applied not only in the selection of organic materials for rechargeable batteries but also in the general prediction of chemical compatibility/stability of a wide range of organic compounds.

Scope of the target reaction and reactivity index with molecular representation

Side reactions are typically induced by the interactions among two or more reactants with chemical incompatibility, and the reaction rates of these irreversible reactions are generally governed by the number of molecules constituting the transition-state complex, the so-called molecularity.[34] Because a transition-state complex constituted by more than two molecules is practically less feasible,[34] we focused on the second-order irreversible reactions between two organic molecules (*i.e.*, molecularity of two at the rate-determining step) in the scope of our study, considering frequently observed side reactions during battery operations and the comprehensively available database for organic materials.[35,36] As an index to determine the reaction reactivity, we selected nucleophilicity and electrophilicity, following the scheme of Mayr *et al.*[37] in determining the rate constant of the reaction between two molecules. According to the scheme, the reaction rate constant of the second-order reactions can be expressed as follows, assuming that extrinsic factors such as mass transport or physical contact are satisfied:

$$\log_{10} k_{20°C} = S_N(N + E).$$

Here, $k_{20°C}$ is the reaction rate constant at 20 °C; S_N is the sensitivity, a coefficient defined by the nucleophile; N is the nucleophilicity of the nucleophile; and E is the electrophilicity of the electrophile. We extracted values of S_N, N, and E from Mayr's database,[37–39] which provides chemical information on a range of organic molecules that have been experimentally verified. Among 1215 of S_N and N values and 317 of E values available in the database, we used 811 of S_N and N values and 217 of E values whose SMILES representations were provided by the database. With 811 values of S_N and N and 217 values of E, we combined S_N, N, and E values to fabricate one datapoint, resulting in a much larger dataset size (175 987). For each datapoint, the reaction rate constant was calculated before training, resulting in a dataset of 175 987 datapoints, consisting of the molecular structures of the nucleophile and electrophile

This journal is © The Royal Society of Chemistry 2023

J. Mater. Chem. A, 2023, **11**, 12784–12792 | **12785**

and the corresponding $\log_{10} k_{20\,°C}$ values. A few examples of the constructed dataset are listed in Table S1.† We used 158 389 datapoints (~90% of the total) as a training/validation set and 17 598 points (~10% of the total) as a test set in our ML scheme.

In each process of the training/validation and test, the molecular structures of the reactants were also considered and were subsequently encoded in a machine-readable form, as we speculated that it would help reveal the local substructure that primarily contributes to the chemical reactivity of the target organic molecule. Here, we adopted a sentence-like embedding method based on the extended connectivity molecular fingerprint (ECFP)[40] to exploit the potential merits of the sentence-like form in applying attention techniques,[41–43] which will be further discussed later. The encoding process was performed in a two-step manner. First, we used canonicalized SMILES (simplified molecular input-line entry system) to represent the molecular structure[44] and converted them into ECFP form. After the conversion process, each molecule structure was expressed by
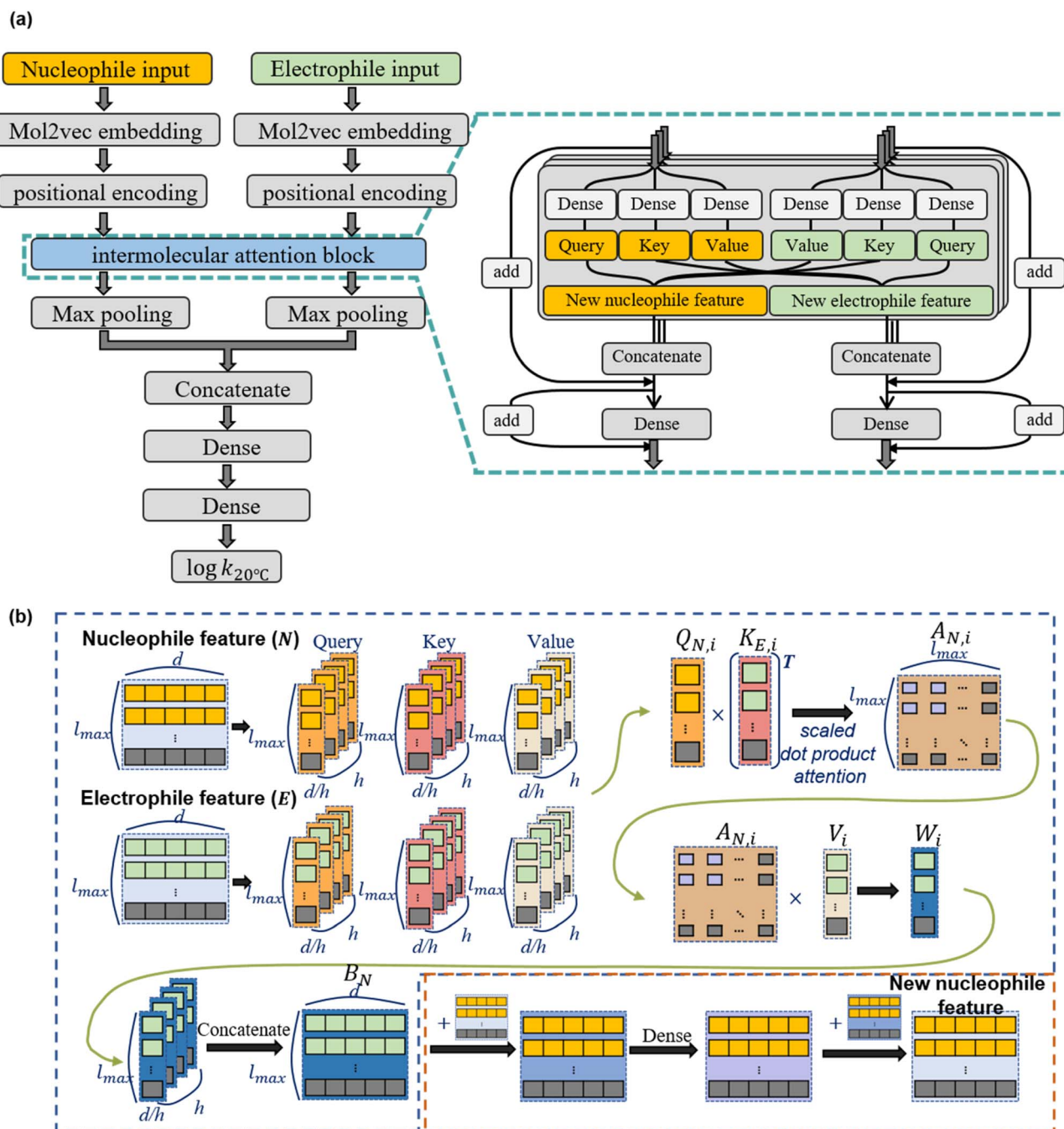

Fig. 1 (a) Architecture of ImRRNet and the working mechanism of intermolecular attention blocks (green dotted line) in ImRRNet. The block includes multi-head attention between two inputs and the following dense layer. Multi-head attention is conducted individually for an individual attention head. (b) Detailed constitution of the intermolecular attention block. The blue and orange boxes indicate the multi-head attention layer and dense layer in the block, respectively.

12786 | J. Mater. Chem. A, 2023, 11, 12784–12792

This journal is © The Royal Society of Chemistry 2023

several integers, and each integer served as an identifier of a substructure in the entire molecular structure. In the second step, we converted the substructure identifiers in the ECFPs into a d-dimensional vector using the pre-trained mol2vec model,[45] which employs word-embedding methods[46] to convert the identifiers into a fixed-size continuous vector. The d-dimensional continuous representation makes it possible to significantly reduce the input dimensionality, which subsequently streamlines the parameter optimization process in the ANN model. Moreover, because embedding vectors contain information concerning the chemical similarity of distinct substructures, training the ANN model becomes much more feasible without additional feature-extracting layers (further details on the molecule representations and the input feature preparation are provided in the Computational method section).

Architecture of ImRRNet, a new ML model

We designed a ML model that predicts the rate constant of a reaction between two arbitrary molecules from their structural data represented by sentence-like forms. The key to the high-quality prediction of our model is the 'attention' technique that helps the ANN model quantitatively evaluate the relative reactivity (*i.e.*, attention weight) of each substructure of molecules. This attention scheme is analogous to what has been practiced in the natural language processing field to swiftly interpret the meaning of a sentence from the attention weights.[47,48] We speculated that ImRRNet can learn and understand how to identify important substructures in inferring reaction rate constants by appraising the mutual attention weights between substructures of molecules during the second-order chemical reaction.[42] ImRRNet consists of an input layer, a positional encoding layer, an intermolecular attention block, a max-pooling layer, and one or more fully connected layers (or dense layers in Tensorflow terminology), as shown in Fig. 1(a). The positional encoding was necessary to maintain the order information because ImRRNet uses a feed-forward neural network instead of a recurrent or convolutional network.[42] The intermolecular attention block is composed of multi-head attention layers and dense layers with residual connection.[49] The block was connected to the two-dimensional max pooling layer and concatenation layer, finally inferring the logarithm of the reaction rate constant.

The intermolecular attention block computes the attention weights of molecular substructures and updates features accentuating the important substructures. Fig. 1(b) illustrates the process of updating the input features of the intermolecular attention block, mainly describing how the nucleophile feature is updated by the block. First, the multi-head attention layer computes the attention scores of each molecular substructure from two input features (*i.e.*, N and E) and updates the input features recursively to obtain attention scores that reflect the contribution of substructures in inferring the reactivity. During the updating process, ImRRNet calculates queries, keys, and values from the input features h times, where h becomes the number of attention heads, forming a 'multi-head' attention

framework. From the computed queries, keys, and values, two attention score matrices can be attained: (i) $A_{N,i}$ from the cross product of $Q_{N,i}$ and $K_{E,i}$ and (ii) $A_{E,i}$ from $Q_{E,i}$ and $K_{N,i}$, such that the obtained attention scores indicate the importance weight matrices of the substructures for the cases in which one molecule 'looks' at the molecule of interest. Here, the attention score matrices, A_i, inform which part of the substructure in the nucleophile (electrophile) molecule has been important in inferring the reaction rate constant with respect to the counter electrophiles (nucleophiles). Attention value matrices are obtained from the attention score matrices and values V_i, which are concatenated to calculate the final features of the attention layer, B_N (or B_E). Further details are provided in the Computational method section.

Attention score of substructures

We can investigate how ImRRNet provides the substructures of significant importance when inferring the chemical reaction rate by analyzing the attention score matrices, *i.e.*, the importance weight matrices. We first trained the ImRRNet model with two heads in the multi-head scheme and evaluated the attention scores for well-known chemical reactions such as the Williamson ether synthesis process for validation. Fig. 2(a) depicts the attention scores inspected for the Williamson ether synthesis process between 1-propanolate and 1-chloropropane, whose main reaction is known to be the nucleophilic attack of 1-propanolate to 1-chloropropane.[50] The trained model suggests that the logarithm of the reaction rate constant is 0.179, which successfully predicts that the reaction will readily occur, consistent with the common experimental knowledge. In Fig. 2(b), the relative attention scores between these two molecules are visualized. Notably, most of the attention scores are concentrated near the O atom in 1-propanolate and near the C atom bonded with the Cl atom in 1-chloropropane, *i.e.*, the reaction sites. This finding is consistent with the general chemical intuition that important substructures that induce chemical reactions are specific sites such as the electron-donating (or withdrawing) group or the leaving group.[50] Furthermore, it demonstrates that our ImRRNet model effectively identifies the reaction sites even without prior chemical knowledge or complex data on the electronic structure of the materials.

We inspected the attention scores more closely from two individual attention heads regarding the relative attention scores and the corresponding submolecular structures. Eight representative molecular substructures are displayed for each reactant molecule along with arrows indicating the particular attention between the substructures in Fig. 2(c). The upper substructures are those of the electrophile 1-chloropropane, and those below are of the nucleophile 1-propanolate. The orange-yellow and green-blue colors are used to visualize the attention from 1-propanolate to 1-chloropropane and *vice versa*, respectively. Among the substructures of 1-chloropropane and 1-propanolate, the **2** and **4** substructures in 1-chloropropane received the greatest attention from 1-propanolate, whereas the **14** and **16** substructures in 1-propanolate received the most

This journal is © The Royal Society of Chemistry 2023

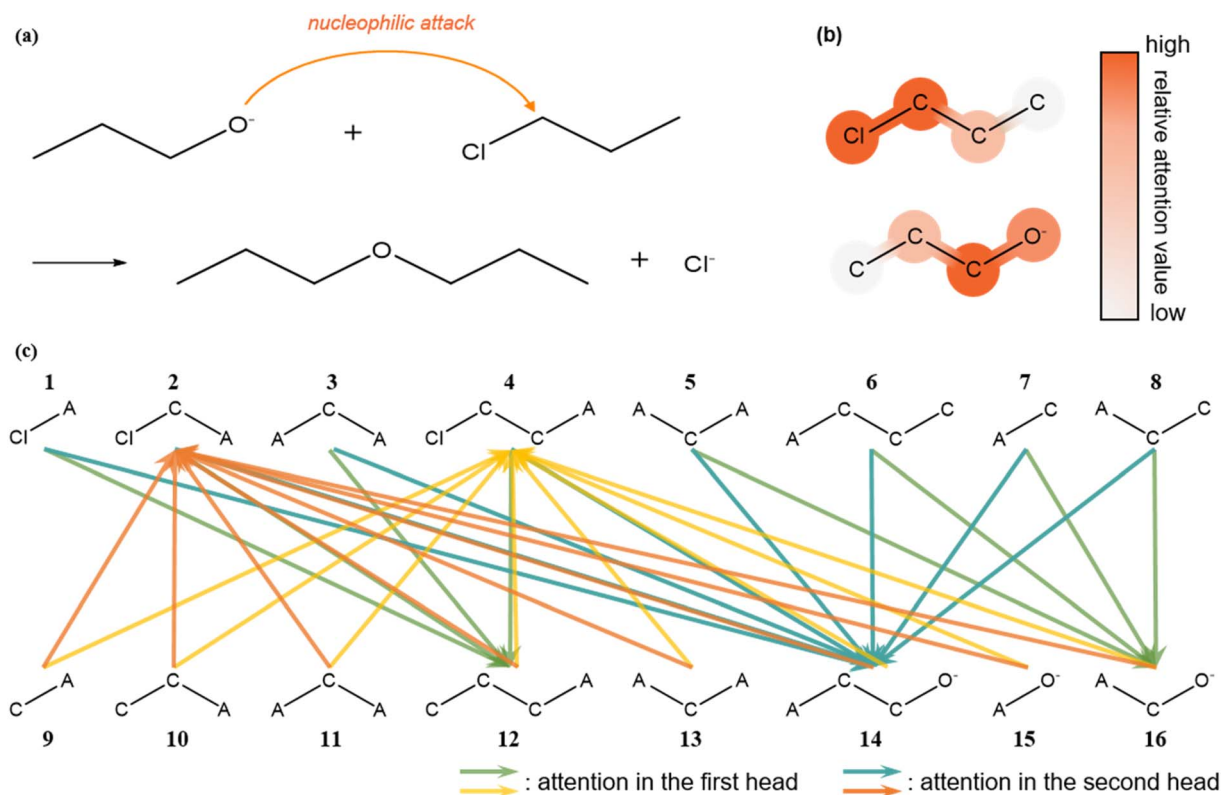*J. Mater. Chem. A*, 2023, **11**, 12784–12792 | **12787**

Fig. 2 (a) Nucleophilic reaction between 1-propanolate and 1-chloropropane during the Williamson ether synthesis process. (b) Relative total amount of attention of 1-propanolate and 1-chloropropane from each other. The relative intensity of the attention is presented using the color bar scheme on the right. (c) Attention given among substructures of 1-propanolate and 1-chloropropane calculated by different attention heads (*e.g.*, the first and second head). Because mol2vec embedding atomic substructures contain information about the number of chemical bonds that each atom forms with non-hydrogen atoms, an arbitrary atom (−A) is introduced to represent information about its chemical bonding. The arrows indicate the substructures that received the most attention from counterpart substructures in one attention head. Orange-like and green-like colors indicate attention given to 1-chloropropane from 1-propanolate and *vice versa*, respectively.

attention from the counter molecule. Notably, we observed that the most reactive substructures are not always mutually paying the greatest attention to each other. Moreover, it was revealed that the attention scores of substructures including neighboring atoms and their bonds (**2**, **4**, **14**, and **16**) were much higher than those of substructures only including a single atom (**1** and **15**). This finding indicates that the chemical reactivities of molecules have relatively little relation to the type of atom itself but are governed by how they bond with each other. It is also noteworthy that relatively high attention scores were recorded for good leaving-group-containing substructures (**2** and **4**). This series of observations clearly indicates that the manner in which ImRRNet predicts the reaction constant is based on the attention mechanism that agrees with common chemical intuition, rather than in the black-box-manner on which most ML techniques rely.

Performance level of ImRRNet

Inspired by the capability of ImRRNet to provide the relevant attention, we comparatively investigated the performance of the ImRRNet model in predicting the reaction-rate constants among any arbitrary combination of organic compounds relative to that of three recurrent neural network (RNN)-based

models that have been previously proposed.[34,51,52] Moreover, we note that a simple linear regression model using several selected molecular properties as input could hardly make accurate prediction of chemical reactivities according to our previous study.[27] The gated recurrent unit (GRU) and long short-term memory (LSTM) models are two of the most commonly used ML techniques that exploit sentence-like data representation similar to ImRRNet. The Delfos model that has been recently introduced is also RNN-based and analogous to the GRU and LSTM models but employs a normal attention technique.[33,43] In particular, Lim *et al.* successfully employed the Delfos model to predict the solvation energy utilizing the attention scheme between organic molecules.[33] In Table 1, the performance of the models is comparatively evaluated with respect to MAE and the R^2 score in predicting the $\log_{10} k_{20°C}$ value, *i.e.*, the reaction rate constant, after hyperparameter tuning for each model. The results clearly indicate that ImRRNet can exhibit significantly better performance (MAE: 0.5760 and R^2 value: 0.9806) than the other models even with a similar number of trainable parameters, suggesting the superiority of the ImRRNet architecture. The Delfos-based model performed slightly better than the GRU and LSTM models, which can be attributed to the employment of the

12788 | *J. Mater. Chem. A*, 2023, **11**, 12784–12792

This journal is © The Royal Society of Chemistry 2023

Table 1 Types of tested artificial neural networks and their attributes/performance. All models were trained with the entire training set with the chosen hyperparameter set

Model	Neural network type	Number of trainable parameters	Attention type	Mean absolute error	R^2 score
GRU	RNN	1 404 401	None	1.2309	0.9589
LSTM	RNN	1 582 801	None	1.0452	0.9668
Delfos-based model	RNN	3 925 601	RNN-based attention	0.9402	0.9703
ImRRNet (this work)	FCNN	1 223 801	Multi-head attention	0.5760	0.9806

attention technique. Nevertheless, its performance was inferior to that of ImRRNet in terms of both the MAE and R^2 values, indicating that the multi-head attention method is more efficient than the RNN-based normal attention method.[42,53] The unique feature of a fully connected neural network (FCNN) of ImRRNet enables the use of a multi-head attention scheme that allows multiple computations of attention values, whereas the Delfos model relies on an RNN-based attention method that calculates the attention value only once. Also, in order to evaluate the general transferability of ImRRNet, we introduced clustering-based cross validation in the reaction rate constant prediction, which evaluates the transferability of the model by rigorously dividing the training set and validation set in terms of molecular similarity (Tables S2 and S3,† see the Computational method for details).[33] When even comparing with previous reports which randomly split the training/validation set, ImRRNet with clustering-base cross validation still showed a considerably lower error in predicting $\log_{10} k_{20°C}$

(MAE of 2.89 and 2.97 for the new-nucleophile test and new-electrophile test, respectively).[27]

Fig. 3(a)–(c) graphically illustrates the prediction results of the ImRRNet, Delfos, and LSTM models, respectively, displaying the predicted $\log_{10} k_{20°C}$ values (y axis) in comparison to the experimental values (x axis). The results indicate that ImRRNet is capable of predicting correct reaction rate constants for a wide range of $\log_{10} k_{20°C}$ values from two arbitrary organic molecules. In contrast, the Delfos and LSTM models tend to overestimate or underestimate the $\log_{10} k_{20°C}$ values with larger MAEs. Notably, when the reaction-rate constants are in the range of exceptionally low (lower-left region of the figure) or high (upper-right region) values, the deviations between the predicted and experimental $\log_{10} k_{20°C}$ values are significantly greater. This finding highlights that the ImRRNet provides sufficient complexity to infer reaction rate constants even for these extreme cases in contrast to the conventional models because of the emphasis on important chemical features in the prediction of the reaction rates. The error distribution
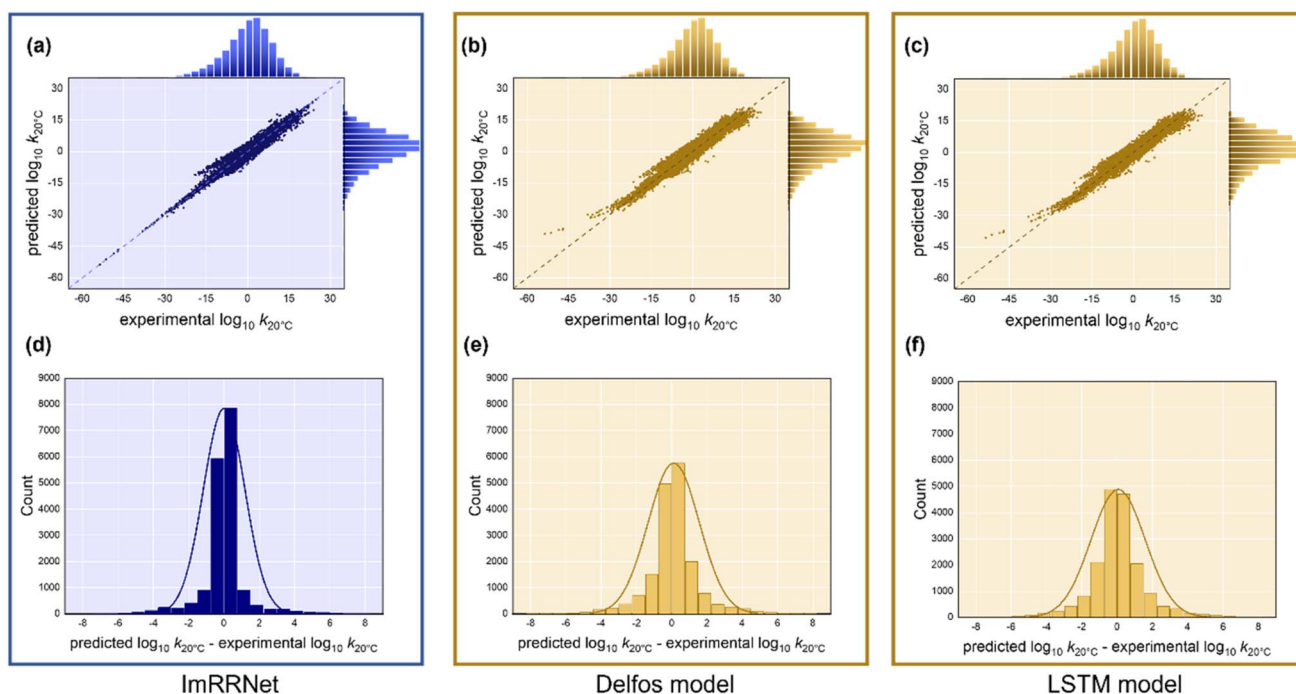

Fig. 3 Performance of (a and d) ImRRNet, (b and e) Delfos-based model, and (c and f) LSTM model when the model was applied to the test set. (a–c) Prediction *vs.* label plot. (d–f) Histogram of error distribution. ImRRNet clearly performs better than other models, especially in predicting low (<−20) and high (>15) $\log_{10} k_{20°C}$ values.

presented in Fig. 3(d–f) also supports the respectable accuracy and reliability of ImRRNet. This figure demonstrates that a much narrower distribution of the error range is guaranteed for the ImRRNet model (Fig. 3(d)) compared with either the Delfos (Fig. 3(e)) or LSTM (Fig. 3(f)) model. Because the range of the error distribution serves as a confidence interval of the prediction, the narrow error distribution is indispensable in providing acceptable reliability, whereas all the ML models naturally contain some extent of intrinsic errors.[27] The standard deviations of the error distributions of the LSTM, Delfos, and ImRRNet models were calculated to be 1.54, 1.30, and 1.05, respectively, which suggests that the 95% confidence interval is ± 3.02, ± 2.55, and ± 2.06 for each model. Using these confidence intervals, we can estimate guideline values of $\log_{10} k_{20°C}$ in screening/selecting chemically stable combinations of organic compounds. For example, if we assume that a value of the rate constant $\log_{10} k_{20°C}$ lower than -6.0 is sufficiently sluggish regarding any side reactions for practical purpose,[37–39] the LSTM, Delfos, and ImRRNet models suggest, with 95% confidence, that the $\log_{10} k_{20°C}$ value of -6.0 lies in the interval of $(-9.02, -2.98)$, $(-8.55, -3.45)$, and $(-3.94, -8.06)$, respectively. This result indicates that the combinations of the two organic materials should have a $\log_{10} k_{20°C}$ value smaller than -9.02, -8.55, and -8.06, respectively, for the materials to be chemically stable with each other, according to the suggestion from each model. Because a greater number of possible combinations can be found with a value of -8.06 using the ImRRNet model, it will be less likely to screen out potentially suitable organic material groups. Moreover, a large uncertain interval is not desirable in the practical screening of materials because the material combinations in the interval should be excluded due to the uncertainty. When applying the LSTM and Delfos-based models, approximately 17.2% and 14.7% of the computed values in the library lie in the uncertain interval, indicating that only the remaining 82.8% and 85.3% of the combinations can be considered in the actual screening of the materials for each model. In contrast, the ImRRNet model offers a smaller uncertain interval of 12.1%, indicating a larger number of candidates (87.9% of the total) for potentially stable organic combinations. The reliable prediction with a smaller error distribution demonstrated here suggests the practical applicability of the ImRRNet model in the screening of various organic compounds and their combinations for stable organic rechargeable batteries.

Summary

We developed a novel neural network model for predicting the rates of chemical reactions between two arbitrary organic chemical species. Combining nucleophilicity and electrophilicity data in Mayr's database, we constructed 175 987 datasets containing organic compounds in a large chemical space, which is one of the most extensive databases. The enriched dataset and multi-head intermolecular attention resulted in outstanding prediction capability of the developed ImRRNet model compared with other RNN-based models and the previous FCNN model[27] on the reaction rate constant $\log_{10} k_{20°C}$.

The accuracy of ImRRNet should be highlighted, as the model was trained with a wide chemical space, which enabled more feasible and accurate prediction of the reaction rate constants for diverse chemicals than previous reports.[32,46] The prediction process used in the ImRRNet model can be strengthened with analysis of the attention scores of individual attention heads, and the predictions were observed to agree well with common chemical intuition. This finding implies that ImRRNet captures the interplay of two sub-molecular structures of reactants and can offer additional knowledge regarding the reaction. We believe that ImRRNet could be widely applicable to various research fields beyond rechargeable organic batteries as well as in the general prediction of the chemical compatibility/stability of a wide range of organic compounds.

Computational method

The Tensorflow 2.2.0 package[54] was employed for implementation of ANN. During hyperparameter optimization of all models, we used the Adam optimizer with a learning rate of 0.0003 and trained for 1300 epochs. After hyperparameter optimization, we re-trained the ImRRNet and other models (GRU/LSTM/Delfos-base) for 1700/1700/2000/2000 epochs with optimized hyperparameters (Tables S4 and S5†). All models were trained by the Adam optimizer whose learning rate started from 0.003/0.003/0.001/0.001 decaying to 0.00015/0.00015/0.00005/0.00005 by sigmoid schedule with the loss function of mean squared error. Different epochs and learning rates were used for a complete convergence. The RDkit package[55] was used for canonicalization of SMILES and for generating ECFPs using the Morgan algorithm. The Morgan algorithm was employed for the conversion process, which generates the identifiers of molecular substructures. Canonicalization of SMILES is necessary for effective training, ensuring unique SMILES representation of a single molecule by reading the substructures in each molecule with a fixed order.[44] Although the information of larger molecular substructures is embedded into the ECFP when the iteration number of the Morgan algorithm increases,[40] only a single iteration of the Morgan algorithm was used for generating the ECFP because the local molecular structure, including the atom itself and the nearest-neighbor atoms, rather than the global structure usually governs chemical reactions. Because the lengths of the ECFPs differed depending on the size of the corresponding molecules, we inserted idle tokens at the end of the generated ECFPs until the length reached the maximum length of the ECFPs in the entire dataset (l_{max}).

The pre-trained mol2vec model was trained with fused data which consisted of the entire dataset from the QM9 database[56,57] and a randomly sampled 10% of the dataset from the ZINC database[58] using the skip-gram algorithm. The mol2vec model embeds the ECFP into a sequence of fixed d-dimensional vectors. Mol2vec maps the molecular substructures into d-dimensional space, locating similar substructures closely by grasping the similarity of the substructures, based on the assumption that similar molecular substructures present similar contexts in the sequence. This assumption is valid because the ECFP contains information of nearby neighbors. In

12790 | *J. Mater. Chem. A*, 2023, **11**, 12784–12792

This journal is © The Royal Society of Chemistry 2023

our scheme, we embedded each molecular substructure into a 300-dimensional vector.

For the preliminary model to explore the attention score between two molecules, we used a small number of heads (2 heads) because it is much more accessible to investigate the attention score of an individual head when using a smaller number of heads. To build an optimal ImRRNet model, we performed 5-fold cross validation. The number of heads, the number of dense layers followed by the attention layer, and the size of the dense layer were set as hyperparameters and optimized using a random search manner. The model with 10 heads and 2 successive dense layers with 400/200 perceptrons was selected as the optimum model. The hyperparameters of reference models (GRU/LSTM/Delfos-base model) were also optimized using a random search manner. To obtain the importance weight matrices of the substructures, we used the scaled dot-product score function for computational simplicity among various score functions.[59] Queries, keys, and values from parameter N in Fig. 1(b) were obtained using the following equations:

$$\text{query}_{N,i}(N) = \boldsymbol{Q}_{N,i} = [\boldsymbol{q}_{N,1}, \boldsymbol{q}_{N,2}, \cdots, \boldsymbol{q}_{N,l_{\max}}]^T$$

$$\text{key}_{N,i}(N) = \boldsymbol{K}_{N,i} = [\boldsymbol{k}_{N,1}, \boldsymbol{k}_{N,2}, \cdots, \boldsymbol{k}_{N,l_{\max}}]^T$$

$$\text{value}_{N,i}(N) = \boldsymbol{V}_{N,i} = [\boldsymbol{v}_{N,1}, \boldsymbol{v}_{N,2}, \cdots, \boldsymbol{v}_{N,l_{\max}}]^T,$$

where $\text{query}_{N,i}$, $\text{key}_{N,i}$, and $\text{value}_{N,i}$ indicate a dense layer applied to calculate the query, key, and value from N, respectively ($i = 1,2,\cdots,h$), and $\boldsymbol{x}_{N,j}$ ($x = q,k,v$ and $j = 1,2,\cdots l_{\max}$, $l_{\max}$ = maximum length of ECFP in the dataset) is a d/h-dimensional intermediate feature vector introduced to compute attention weights. The queries $\boldsymbol{Q}_{E,i}$, keys $\boldsymbol{K}_{E,i}$, and values $\boldsymbol{V}_{E,i}$ were similarly calculated from E. By combining the attention score matrices and values, the attention values were obtained as follows:

$$W_{N,i} = \text{Attention}(\boldsymbol{Q}_{N,i}, \boldsymbol{K}_{E,i}, \boldsymbol{V}_{E,i})$$

$$W_{E,i} = \text{Attention}(\boldsymbol{Q}_{E,i}, \boldsymbol{K}_{N,i}, \boldsymbol{E}_{N,i})$$

$$\text{Attention}(\boldsymbol{Q}_{X,i}, \boldsymbol{K}_{Y,i}, \boldsymbol{V}_{X,i}) = \text{softmax}\left(\frac{\boldsymbol{Q}_{X,i}\boldsymbol{K}_{Y,i}^T}{d/h}\right)\boldsymbol{V}_{Y,i}$$

$$= \boldsymbol{A}_{N,i}\boldsymbol{V}_{Y,i} = \boldsymbol{W}_{X,i}$$

Here, the subscripts X and Y indicate nucleophile (electrophile) and its counterpart molecule, respectively; $\boldsymbol{W}_{X,i}$ is the attention value matrix; and $\boldsymbol{A}_{X,i}$ is the attention score matrix ($(X,Y) = (N,E),(E,N)$). To avoid calculating attention scores for padded features, which were added to match the data length, we introduced the masking technique. In the intermolecular attention block, residual connections were introduced to accomplish faster and finer convergence, which connected the front and back of each attention layer and dense layer, adding exactly the same features of the previous layer to the current layer.

To evaluate the transferability of ImRRNet to new molecular substructures and molecules, K-means clustering base cross validation is introduced. Because K-means clustering splits the dataset by minimizing the variance of embedding vectors within each cluster, each cluster consists of structurally similar molecules. In other words, this means that structures of molecules in different clusters constructed by K-means clustering would be much different. Therefore, after grouping the entire dataset into K groups with K-means clustering, a general capability of the model to predict reaction rate constants can be validated when totally "new" molecules are inputted in the model that has been trained by the dataset of K-1 groups. After embedding molecular structures as 1024-bitwise molecular fingerprints using the RDkit package,[55] we split nucleophile and electrophile structure datasets into 8 groups using K-means clustering, respectively. Two new-molecule tests (inputting new nucleophiles and new electrophiles to the model) were conducted. When testing the model with new nucleophiles (new nucleophile test), the training set consists of one group of nucleophiles and the entire electrophile dataset, whereas the validation set consists of the remaining 7 groups of nucleophiles and the entire electrophile dataset. Alternating the nucleophile group used in the training set, we trained the model 8 times. For the case of the new electrophile test, an analogous methodology was used.

The final model and code of ImRRNet are available at https://github.com/yanselmo/ImRRNet-Inter-MolecularReactionRateNetwork-.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

References

1 Y. Zhao, O. Pohl, A. I. Bhatt, G. E. Collis, P. J. Mahon, T. Rüther and A. F. Hollenkamp, *Sustainable Chem.*, 2021, **2**, 167–205.

2 C. Curry, *Bloomberg New Energy Finance*, 2017, vol. 5, pp. 4–6.

3 L. Wang, J. Hu, Y. Yu, K. Huang and Y. Hu, *J. Cleaner Prod.*, 2020, **276**, 124244.

4 S. Kiemel, S. Glöser-Chahoud, L. Waltersmann, M. Schutzbach, A. Sauer and R. Miehe, *Resources*, 2021, **10**, 84.

5 S. Lee, G. Kwon, K. Ku, K. Yoon, S.-K. Jung, H.-D. Lim and K. Kang, *Adv. Mater.*, 2018, **30**, 1704682.

6 Y. Lu and J. Chen, *Nat. Rev. Chem.*, 2020, **4**, 127–142.

7 S. Lee, J. Hong and K. Kang, *Adv. Energy Mater.*, 2020, **10**, 2001445.

This journal is © The Royal Society of Chemistry 2023

J. Mater. Chem. A, 2023, **11**, 12784–12792 | **12791**

8 J. Winsberg, T. Hagemann, T. Janoschka, M. D. Hager and U. S. Schubert, *Angew. Chem., Int. Ed.*, 2017, **56**, 686–711.

9 J. Luo, B. Hu, M. Hu, Y. Zhao and T. L. Liu, *ACS Energy Lett.*, 2019, **4**, 2220–2240.

10 M. Lee, J. Hong, B. Lee, K. Ku, S. Lee, C. B. Park and K. Kang, *Green Chem.*, 2017, **19**, 2980–2985.

11 H. W. Kim, H.-J. Kim, H. Byeon, J. Kim, J. W. Yang, Y. Kim and J.-K. Kim, *J. Mater. Chem. A*, 2020, **8**, 17980–17986.

12 D. G. Kwabi, Y. Ji and M. J. Aziz, *Chem. Rev.*, 2020, **120**, 6467–6489.

13 H. Fan, W. Wu, M. Ravivarma, H. Li, B. Hu, J. Lei, Y. Feng, X. Sun, J. Song and T. L. Liu, *Adv. Funct. Mater.*, 2022, 2203032.

14 H. Fan, B. Hu, H. Li, M. Ravivarma, Y. Feng and J. Song, *Angew. Chem., Int. Ed.*, 2022, **61**, e202115908.

15 X. Wei, W. Xu, J. Huang, L. Zhang, E. Walter, C. Lawrence, M. Vijayakumar, W. A. Henderson, T. Liu, L. Cosimbescu, B. Li, V. Sprenkle and W. Wang, *Angew. Chem., Int. Ed.*, 2015, **54**, 8684–8687.

16 J. Back, G. Kwon, J. E. Byeon, H. Song, K. Kang and E. Lee, *ACS Appl. Mater. Interfaces*, 2020, **12**, 37338–37345.

17 G. Kwon, K. Lee, J. Yoo, S. Lee, J. Kim, Y. Kim, J. E. Kwon, S. Y. Park and K. Kang, *Energy Storage Mater.*, 2021, **42**, 185–192.

18 K. Jorner, A. Tomberg, C. Bauer, C. Sköld and P.-O. Norrby, *Nat. Rev. Chem.*, 2021, **5**, 240–255.

19 H. Chermette, *J. Comput. Chem.*, 1999, **20**, 129–154.

20 H. Eyring, *J. Chem. Phys.*, 1935, **3**, 107–115.

21 H. Park, H.-D. Lim, H.-K. Lim, W. M. Seong, S. Moon, Y. Ko, B. Lee, Y. Bae, H. Kim and K. Kang, *Nat. Commun.*, 2017, **8**, 14989.

22 L. R. Domingo, M. Ríos-Gutiérrez and P. Pérez, *Molecules*, 2016, **21**, 748.

23 R. G. Parr, R. A. Donnelly, M. Levy and W. E. Palke, *J. Chem. Phys.*, 1978, **68**, 3801–3807.

24 C. L. Haynes, Y.-M. Chen and P. B. Armentrout, *J. Phys. Chem.*, 1995, **99**, 9110–9117.

25 Y. Guan, V. M. Ingman, B. J. Rooks and S. E. Wheeler, *J. Chem. Theory Comput.*, 2018, **14**, 5249–5261.

26 L. R. Domingo and P. Pérez, *Org. Biomol. Chem.*, 2011, **9**, 7168–7175.

27 B. Lee, J. Yoo and K. Kang, *Chem. Sci.*, 2020, **11**, 7813–7822.

28 F. Palazzesi, M. R. Hermann, M. A. Grundl, A. Pautsch, D. Seeliger, C. S. Tautermann and A. Weber, *J. Chem. Inf. Model.*, 2020, **60**, 2915–2923.

29 C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay and K. F. Jensen, *Chem. Sci.*, 2019, **10**, 370–377.

30 S. Zhong, K. Zhang, D. Wang and H. Zhang, *Chem. Eng. J.*, 2021, **405**, 126627.

31 G. V. S. M. Carrera, S. Gupta and J. Aires-de-Sousa, *J. Comput.-Aided Mol. Des.*, 2009, **23**, 419–429.

32 M. Orlandi, M. Escudero-Casao and G. Licini, *J. Org. Chem.*, 2021, **86**, 3555–3564.

33 H. Lim and Y. Jung, *Chem. Sci.*, 2019, **10**, 8306–8315.

34 L. Arnaut and H. Burrows, *Chemical Kinetics: from Molecular Structure to Chemical Reactivity*, Elsevier, 2006.

35 H. E. Pence and A. Williams, *J. Chem. Educ.*, 2010, **87**, 1123–1124.

36 S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He and B. A. Shoemaker, *Nucleic Acids Res.*, 2016, **44**, D1202–D1213.

37 H. Mayr, *Tetrahedron*, 2015, **32**, 5095–5111.

38 Z. Li, H. Jangra, Q. Chen, P. Mayer, A. R. Ofial, H. Zipse and H. Mayr, *J. Am. Chem. Soc.*, 2018, **140**, 5500–5515.

39 R. J. Mayer, M. Breugst, N. Hampel, A. R. Ofial and H. Mayr, *J. Org. Chem.*, 2019, **84**, 8837–8858.

40 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.

41 C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, *Chem. Mater.*, 2019, **31**, 3564–3572.

42 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, *Advances in Neural Information Processing Systems*, 2017.

43 D. Bahdanau, K. Cho and Y. Bengio, *arXiv*, 2014, preprint arXiv:1409.0473.

44 M. Hirohara, Y. Saito, Y. Koda, K. Sato and Y. Sakakibara, *BMC Bioinf.*, 2018, **19**, 83–94.

45 S. Jaeger, S. Fulle and S. Turk, *J. Chem. Inf. Model.*, 2018, **58**, 27–35.

46 T. Mikolov, K. Chen, G. Corrado and J. Dean, *arXiv*, 2013, preprint arXiv:1301.3781.

47 A. Trewartha, N. Walker, H. Huo, S. Lee, K. Cruse, J. Dagdelen, A. Dunn, K. A. Persson, G. Ceder and A. Jain, *Patterns*, 2022, **3**, 100488.

48 L. Li, J. Wan, J. Zheng and J. Wang, *BMC Bioinf.*, 2018, **19**, 285.

49 K. He, X. Zhang, S. Ren and J. Sun, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

50 J. S. Gorzynski, *Organic Chemistry*, McGraw-Hill/Higher Education, 5th edn, 2016.

51 S. Hochreiter and J. Schmidhuber, *Neural Comput.*, 1997, **9**, 1735–1780.

52 K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, *arXiv*, 2014, preprint, arXiv:1406.1078.

53 J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, *arXiv*, 2018, preprint arXiv:1810.04805.

54 M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean and M. Devin, *arXiv*, 2016, preprint arXiv:1603.04467.

55 *RDKit: Open-Source Cheminformatics*, https://www.rdkit.org, DOI: 10.5281/zenodo.3815117.

56 L. Ruddigkeit, R. Van Deursen, L. C. Blum and J.-L. Reymond, *J. Chem. Inf. Model.*, 2012, **52**, 2864–2875.

57 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. Von Lilienfeld, *Sci. Data*, 2014, **1**, 1–7.

58 T. Sterling and J. J. Irwin, *J. Chem. Inf. Model.*, 2015, **55**, 2324–2337.

59 Y. Shen, E. M. K. Lai and M. Mohaghegh, *Neural Process. Lett.*, 2022, **54**, 2283–2302.

12792 | *J. Mater. Chem. A*, 2023, **11**, 12784–12792

This journal is © The Royal Society of Chemistry 2023