THE JOURNAL OF PHYSICAL CHEMISTRY B

pubs.acs.org/JPCB Article

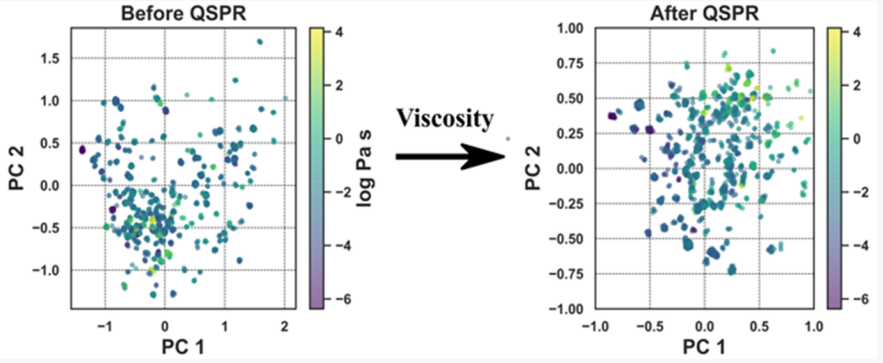
Continuous Molecular Representations of Ionic Liquids

Published as part of The Journal of Physical Chemistry virtual special issue "Machine Learning in Physical Chemistry".

Wesley Beckner,[†] Chowdhury Ashraf,[†] James Lee, David A. C. Beck, and Jim Pfaendtner*

Cite This: *J. Phys. Chem. B* 2020, 124, 8347–8357 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Designing new ionic liquids (ILs) is of crucial importance for various industrial applications. However, this always leads to a daunting challenge, as the number of possible combinations of cation and anion are very high and it is impossible to experimentally propose and screen a wide pool of potential candidates. However, recent applications of machine learning (ML) models have greatly improved the overall chemical discovery pipeline. In this study, we compare different generative methods for producing ionic liquids. In this comparison, we show the following: (1) when training data is scarce, a transfer learning approach can be applied to variational autoencoders (VAEs) to generate molecular structures of the target molecule type; (2) in a VAE-like structure, separate latent spaces for the cationic and anionic moieties can result in meaningful representations for their combinative, macroscopic properties; (3) interpolating between ILs with desired properties can result in a new IL with attributes similar to the two structural end points.

INTRODUCTION

Applications of ionic liquids (ILs) are experiencing a recent surge in various industries due to their versatile properties. For example, properties like low flammability, negligible vapor pressure, high thermal stability, and wide electrochemical window have made them suitable for energy storage applications.[1] Other applications of ILs include but are not limited to solvents for pharmaceuticals,[2] CO_2 capture,[3] catalysis and biocatalysis,[4,5] cellulose dissolution,[6] and organic synthesis.[7] However, with theoretically possible 10^{14}–10^{18} molecular configurations,[8] experimental optimization of potential ILs for targeted applications is a daunting task, since their specific properties largely depend on the structure, dynamics, and interaction of their constituent anions and cations. In recent years, physics based simulations such as molecular dynamics (MD) and Monte Carlo (MC) have shown promise to supplement wet lab experiments.[9] While MD/MC simulations have been able to accurately predict some of the properties of ionic liquids such as density,[10] due to high computational cost associated with these simulations and a need for accurate interatomic potentials, they are useful only for validating a few selected candidates, not for screening a vast pool of possible candidates.[11] Therefore, more refined methodologies are required not only to predict properties without doing expensive time-consuming experiments but also to screen potential candidates from available databases, and in some cases, propose new candidates with desired properties. To this end, recent advances in data science and machine learning techniques are playing a pivotal role in high throughput screening and analysis of a large number of IL samples.[12−16]

In general, machine learning models, particularly deep learning ones (typically categorized as neural networks

Received: June 29, 2020
Revised: August 21, 2020
Published: August 24, 2020

https://dx.doi.org/10.1021/acs.jpcb.0c05938
J. Phys. Chem. B 2020, 124, 8347–8357

(NNs) with three or more layers), are significantly accelerating the overall chemical discovery pipeline. These include highly accurate quantitative structure−property relationships (QSPRs) for predicting properties of candidate molecules, generative modeling to propose new candidate molecules for specific applications, machine learning accelerated simulations, and machine learned, reliable interatomic potentials.[17−21] Since 2012, advances in GPU-accelerated training and regularization tricks like dropout have put deep learning at the forefront of the molecular design toolkit.[22] This accelerated training comes on the heels of back-propagation, the algorithm by which connectionist-type learners can appropriate blame in their network weights and therefore achieve gradient based solutions to mastering their training data. Indeed, the similarity of weighting interconnected neural layers to traditional graphical processing tasks has led to the creation of specialized hardware for those purposes.[23] Further, modular Python libraries like Keras[24] and TensorFlow[25] have made appropriation of deep learning in computational molecular science very convenient for a range of creative applications. Lastly, the ability of deep learners to embed discretized objects into a continuous space further opens up the design paradigm into rapid, gradient-based solutions.

In this study, we aim to develop a generative deep learning model for designing new IL molecules. However, this presents a twofold challenge: at one side, deep learning models often contain hundreds of thousands, if not millions, of parameters; i.e., they are extremely data hungry.[26] On the other side, there is a serious lack of large curated experimental data sets that are publicly available for ILs. Even MD/MC simulations cannot come to the rescue due to their previously mentioned limitations. In the context of training such generative models, these data scarcity issues lead to low population of the latent space from which candidate structures are generated and properties predicted. The lack of training examples means that the network may not be able to accurately scale to new types of ionic liquids. However, recent works have investigated how transfer learning or analogical learning might be better utilized in these types of networks to overcome this challenge.[27−32] For example, Gómez-Bombarelli et al. developed a model comprised of a variational autoencoder (VAE) trained simultaneously with a neural network to both generate and predict the property value of drug-like molecules.[33] Additionally, Goh et al. developed ChemNet, a deep neural network (DNN) first pretrained on molecular descriptor labels, under weak supervised learning, and then trained, using transfer learning, on smaller data sets to predict molecular properties.[34] However, transfer learning can suffer from "amnesia", i.e., forgetting generalizations learned in the initial phases of training.[35] To address these challenges in this study, we set out to show that a small molecule database, GDB-17,[36] can be leveraged in a VAE[37−39] to embed cations and anions that can then be used to generate candidate IL structures. A particular challenge with this approach is the choice of training schedule for these deep networks. In the first section of the study, we explore novel training schedules and the balance of retaining general chemical knowledge with the need for IL specific information content.

In generative modeling, molecules should be represented in a suitable way so that a decision or a design criterion can be met. Recently, great progress has been made in representing molecules as molecular graphs, where atoms and bonds make up nodes and edges in a connectivity matrix.[40,41] The main advantages of this approach are that representations are invertible, i.e., every graph is associated with a single molecule, and unique, i.e., every molecule is associated with a single graph. While this may reduce the computational load to generate a unique molecular candidate, it is not necessary for the generator to learn "true" chemistry. Indeed, some strategies for molecular embedding have even been oriented around tasking the generator with learning the relationships between various types of molecular representations.[42] In this work, we have the added challenge that each IL "material" is constituted as a pair of individual yet strongly coupled ionic species—the cation and the anion. Herein, we explore how these two distinct moieties can be represented in a latent space and use SMILES annotation to represent their individual structures.

SMILES, developed by Weininger[43] and Daylight Chemical Information Systems, is an ASCII character string representation of a depth-first search of a molecule's graph. In SMILES representation, a molecule may be invertible but nonunique if it is physically asymmetrical by translation or rotation or if it varies with permutation of atomic indexes. When represented in 2D, such a molecule may have different representations (nonunique). All of these representations, however, still refer to the same molecule (invertible). A nonunique string data set can be made unique by the process of canonicalization, which determines which of all possible string representations of a molecule will be used as the reference for its molecular graph. Certain canonical string representations such as IUPAC's InChI[44] exist, and while canonicalization algorithms can be applied to SMILES,[45] no standard method pervades. This being said, SMILES is the native encoding for many large databases[21] such as ZINC,[46] ChEMBL,[47] and GDB-17 which we leverage in this work; therefore, SMILES was selected as the molecular representation for the machine learning tasks herein. As it is a string sequence, we can take advantage of sequence-based deep learning models such as VAEs, which have found success in natural language processing (NLP) and recently in molecular generative models for drug-like molecules.[33,48−50] By that same token, the SMILES string is fragile, meaning a small change in syntax can result in a chemically infeasible molecule. For this reason, a method of guiding the search through chemical space is required. To validate our VAE outputs, we utilize a sanitization or "chemical feasibility" step from RDKit[51] to check the atomic valences of emergent structures.

A goal of a generative model is to explore the chemical space while achieving optimization of a single molecular property or multiobjective optimization of a set of molecular properties. In order to traverse the entirety of small molecule space (the largest training cation had 37 heavy atoms, and ion SMILES strings were limited to under 62 characters) and optimize over it, a continuous, gradient-based generative model is desired. This makes VAE an attractive option for the chemical design of IL materials. A VAE is comprised of two neural networks: an encoder and a decoder. The input to the encoder is a one-hot encoded SMILES string, a common method of vector representation that avoids unwanted mathematical relationships between inputs due to orthogonality between representative vectors. In our study, the encoder converts one-hot encoded SMILES strings into a latent vector representation, and the decoder converts the encoded string back into a SMILES string. The output of the encoder and input of the decoder are of low dimension compared to the dimensionality of the hidden layers of the encoder and decoder. For this

https://dx.doi.org/10.1021/acs.jpcb.0c05938
J. Phys. Chem. B 2020, 124, 8347−8357

reason, it is known as a bottleneck layer. In training, the model must learn to represent data as best as possible in this bottleneck layer, learning some representation of the key features of a molecule. This is not unlike the process of dimensionality reduction in principal component analysis. The vector representation of the encoded string is then known as its latent representation, a single distribution in the latent space. By using the latent representation of a string as the input to a QSPR neural network and training the entire model to include the loss from the QSPR predictor, one can organize the latent space in relation to both structure and property.[33] The latter sections of this work, therefore, explore the utility of the latent by generating new candidate IL molecules from the latent space and investigating their properties.

In general, the goals of this study are to establish a transfer learning protocol with VAEs to build a generative model for IL materials where experimental data is scarce and then add a predictive model with the generative one to propose new IL structures with desired properties. The rest of this paper is organized as follows: in the next section, we elaborated on the working principle of variational autoencoders (VAEs) and discussed various generative models for producing ionic liquids and the rationale behind choosing them to compare in this study. Then, in the Results and Discussion, we established the best model for our generative learning scheme, compared various network architectures to handle the latent space of anions and cations, and selected the best performing one. Next, we demonstrated that it is possible to generate novel IL candidates with desired properties by interpolating between two different ILs. Lastly, we concluded with some remarks and future directions.

METHODS

The Variational Autoencoder. The variational autoencoder (VAE) has seen success in the generation of images[52] and has also been recently used for the generation of drug-like molecules.[33] It was therefore an attractive model for the chemical design of ionic liquid materials. A variational autoencoder is an autoencoder with added stochasticity. An autoencoder is comprised of two neural networks: an encoder and a decoder. The encoder inputs data input x and outputs a latent (hidden) representation, which is of much lower dimensionality than x. The decoder takes the latent representation z (a vector) and outputs a prediction $\hat{x}$. In training, both the data and target label are the input x, so that the model learns to predict its input. The significance lies in the low dimensionality of the latent representation z, which learns the most important features of the input in order to make accurate predictions. The compression of x to z is like the process of dimensionality reduction in principal component analysis, wherein similar data inputs will be located close together in latent space. In a variational autoencoder, the input x is no longer mapped to a single vector in latent space. Instead, it is mapped to a distribution over latent space. This distribution is the values of z from which x could have been generated. This takes the form of a vector of means and a vector of standard deviations, wherein the ith elements refer to the mean and standard deviation of the distribution corresponding to the ith data point. The decoder samples from this distribution a value z and outputs a distribution of values of x to which z corresponds.[39,53]

In this work, we use the variational autoencoder to generate molecular structures. Just as in a traditional VAE, the objective function includes a reconstruction loss (the ability of the decoder to recreate the input) and a Kullback−Leibler divergence or KL divergence loss (the compactness of encoded distributions).[39] Since we are also interested in the properties of these generated structures, we add to this objective function the loss from the QSPR predictor. This led to a densely populated latent space reorganization in relation to both structure and property. The importance of latent space organization lies in sampling new structures. In sampling from the latent space, samples closely related in terms of structure and chemical property will be close together in latent space. If one desires to generate a structure with high density, one could then use the latent representation of a known molecule with high density, specify a temperature (meaning a distance from that point), and begin sampling. If one desires to generate molecules with structure and property between two known molecules, one can sample from the interpolation between these two points in latent space. In this work, we use spherical linear interpolation (SLERP),[54] which is discussed in detail later.

Transfer Learning Protocol. In this study, the salt data were taken from the ILThermo[55,56] database with experimental entries for properties: density, heat capacity, viscosity, and thermal conductivity. This database consisted of 688 unique entries comprised of 276 unique cations and 98 unique anions. This is a challenging example of a scarce data set where canonical machine learning approaches cannot be used with a high degree of confidence. As stated previously, researchers are using transfer learning approaches to overcome the challenge of data scarcity. In a transfer learning protocol, the model initially learns from a large data set and applies the stored information to learn from a different but related and possibly scarce data set. To prove the concept that transfer learning works for our problem and to establish a protocol, we investigated five different protocols (M1−M5, Table 1)—three non-transfer learning (purist) and two transfer learning protocols. All of the protocols were instantiated using a VAE structure and hyperparameters developed previously for small, drug-like molecules.[33] As this was a procedure to establish a transfer learning protocol, only the molecules of the cationic moiety were targeted. In later sections when attempting to create ILs with specific properties, both the cation and anion were embedded.

Table 1. VAE Training Protocols for Models M1−M5[a]

model	GDB-17 training samples	1:1 GDB-17:bootstrapped cation training samples (N = 276)	bootstrapped cation training samples (N = 276)
M1			250,000
M2			1,000,000
M3	1,000,000		
M4	1,000,000	500,000	
M5	1,000,000	500,000	500,000

[a]Only the cationic moieties are considered.

The first three models (purist models, M1−M3) are trained either on the GDB7−17 database or on the cation database (ILThermo, 276 unique structures). GDB-17 is a well-known database for small organic molecules with roughly 166 billion structures and is very much suitable for training machine learning models. Conversely, the cation database is much smaller and therefore bootstrapping was employed. Boot-

https://dx.doi.org/10.1021/acs.jpcb.0c05938
J. Phys. Chem. B 2020, 124, 8347−8357

strapping is a powerful tool which quantifies the uncertainties associated with any measured data set by increasing the number of data points through sampling with replacement, and this bootstrapped ensemble modeling approach has been used previously in other application areas such as hydrological forecasting.[57−59] By bootstrapping the cation data set, we generated 250,000 and 1,000,000 training samples in models M1 and M2, respectively, while, in M3, we randomly selected 1,000,000 training samples from the large GDB-17 data set. These purist models are generated as controls to the transfer learning protocols. Next, we perform two additional transfer learning protocols (models M4 and M5). In both transfer learning protocols, all weights could be updated in every training iteration. In model M4, we start from model M3 which has already learned from the GDB-17 data set and trained it on a different data set to train a new model. This new data set contains 500,000 training samples, but this time, 250,000 of them are randomly selected from the previously sampled GDB-17 data set of 1,000,000 samples and the remaining 250,000 are bootstrapped from the cation data set. Lastly, our final transfer learning protocol, model M5, is built starting from model M4 and then trained on a data set of 500,000 training samples generated only by bootstrapping the cation data set. It may initially seem nonintuitive to bootstrap such large sample sets from the comparatively small cation sample size. Indeed, the purpose of the M1 and M2 models is to demonstrate the folly of exposing a model to too many rounds or epochs of the same data—a practice that results in overfitting. What will be shown in the transfer learning approaches is that the effect of an initialized, nonrandom weight distribution, that is, a neural architecture that has been "pre-fabricated" according to a large variety of chemical structures, will be to enable these architectures to learn from the bootstrapped data sets without sacrificing their ability to create structurally diverse chemicals. This will be demonstrated by the ability of these resultant models (M4 and M5) to generate structurally diverse molecules from latent cation seeds.

RESULTS AND DISCUSSION

Transfer Learning Approach. To evaluate the effectiveness of our transfer learning protocol, we created five models (three purist models and two transfer learning models) using identical architectures and training data yet different training protocols. The training protocols are listed in Table 1. Once the models were trained, they were tasked with generating structures from the latent space using a randomly selected cation latent seed: 1-butyl-2-methylpyridinium (1 of the 276 cations in the cation data set). If after 10,000 sample attempts the model was unable to procure a new and unique SMILES structure that followed basic chemical rules such as valency—something we will refer to as chemically feasible—the search was terminated. The total number of returned structures from each model is listed in Table 2. Additionally, since our target is to generate cation molecules, we added a charge criterion to generate molecules with positive charge in our molecule generation scheme for the same cation seed. The total number of samples with positive charge returned from each model is also listed in Table 2 as well as indicated in Figure 1, top panel. We note that the only model to not include 1-butyl-2-methylpyridinium in its training data is M3. From Table 2, we can see that model M3 generates the least number of chemically feasible structures; at the same time, this model fails to generate any sample with positive charge. This is expected, as this model is entirely trained on the GDB-17 database which does not have any molecules with positive charge; therefore, the model fails to learn the features associated with positively charged molecules. Model M1 generates the greatest number of chemically feasible structures compared to all other models, but the number of positively charged samples is small even when it is compared to the other purist model M2, which is also trained only on the bootstrapped cation database. When we consider our transfer learning models M4 and M5, we observe that M5 procures the second highest number of chemically feasible structures for both regular samples and positively charged samples among all five models, while the samples produced by model M4 are very low for both cases. This is due to the fact that model M4 was mostly trained on a larger portion of GDB-17 data but tasked with generating cationic samples. Additionally, we compared the chemical (Tanimoto) similarity of the generated structures with the seed structure when only the positively charged samples are generated the using molecular access system (MACCS)[60] fingerprints, as shown in Figure 1, bottom panel. Here, the Tanimoto similarity metric is defined as the

Table 2. Total Samples Generated from Each Model with Single Cation Seed, 1-Butyl-2-methylpyridinium

model	samples	samples with (+) charge
M1	7122	6
M2	170	167
M3	6	0
M4	43	29
M5	3700	1729

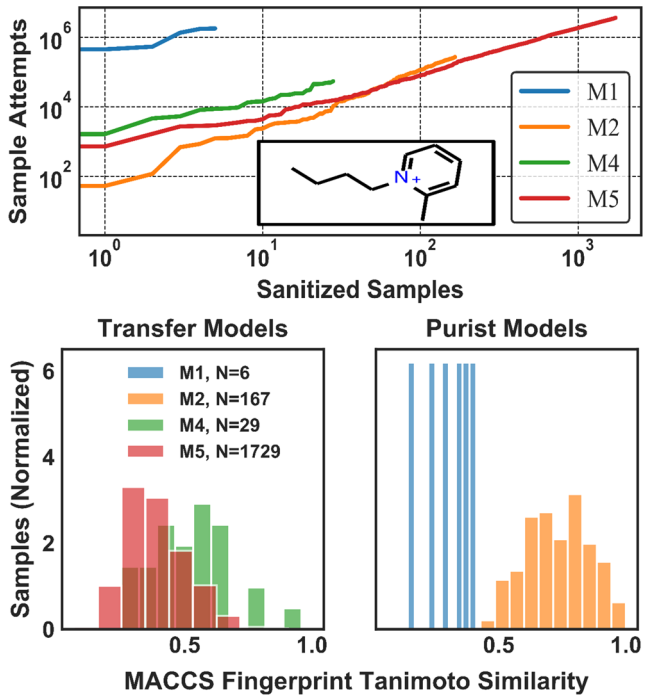


Figure 1. (top) Log−log scale sample attempts vs number of chemically feasible structures with a positive charge (also indicated by *N* in bottom, left panel). Sampled from a single latent space cation seed (inset). (bottom) Tanimoto similarities of MACCS fingerprints of procured structures compared to cationic seed. Lower values are more dissimilar from the seed, and broader histogram distributions contain more structural variety.

https://dx.doi.org/10.1021/acs.jpcb.0c05938
J. Phys. Chem. B 2020, 124, 8347−8357

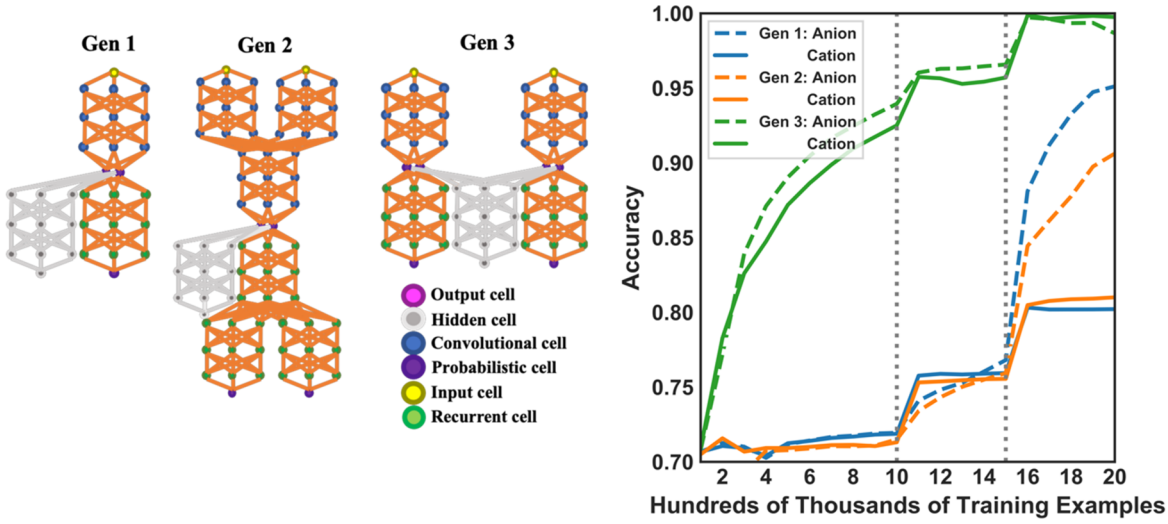


Figure 2. (left) Network architectures of three different VAE models. (right) Historical training accuracy for models Gen1, Gen2, and Gen3. Note: These are the built in categorical accuracy metrics within Keras. They are not the Tanimoto similarities or the loss criteria. They are a per-SMILES-character frequency term indicating how often the output recreates the input to the model. Training protocols are the same as M5. Training accuracies for Gen1 and Gen2 did not appreciably improve with 1 million GDB-17 examples.

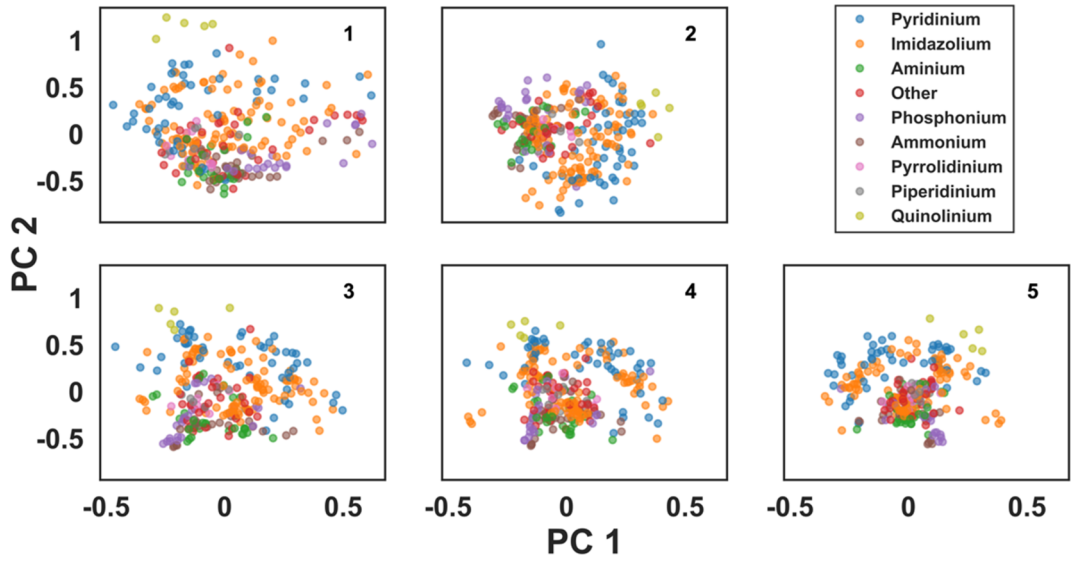


Figure 3. First two principal components during phase III salt embeddings at every 100,000 training examples. The inset number indicates the 100,000th training step.

intersection over the union of the two fingerprints, and this figure clearly demonstrates that M5 produces structures with greater MACCS variety than its companion transfer learned model, M4. The only other model that produces structures with somewhat better MACCS variety is model M2. However, the number of chemically feasible, positively charged structures generated by M2 is an order of magnitude smaller than that for M5. Therefore, M5 clearly achieves the best performance for generating cationic candidates for ILs. This bolsters the fact that the transfer learning protocol can be used in generating novel cationic structures where experimental data is scarce. The transfer learned training protocol for M5 was therefore identified as the training protocol to create dual cation−anion generating VAEs.

Sampling from Dual Latent Spaces for Target Properties. After selecting M5 as having the preferred training protocol, the network architecture was optimized to represent two distinct moieties of ILs, namely, cation and anion, in the latent space. We considered three dual cation−anion VAEs, in order to procure latent space(s) with which to feed a QSPR model, shown in Figure 2, left panel. As a null hypothesis, Gen1 consisted of the Gómez-Bombarelli[33] architecture, albeit with the first and final layers receiving two, and outputting two, SMILES strings for the cation and anion, respectively. Gen2 contained the same structure as Gen1 with the exception that the cation−anion inputs fed into three independent convolutional (CONV) layers before feeding into three combined CONVs and the third gated recurrent unit (GRU) in the decoder fed into two separate branches of three GRUs for each of the cation and anion. Gen3 VAE consisted of two separate Gen1 architectures. Training protocols for the three architectures followed the same protocol as M5. All three models consisted of the same QSPR structure described in previous work,[61] with the

https://dx.doi.org/10.1021/acs.jpcb.0c05938
J. Phys. Chem. B 2020, 124, 8347−8357

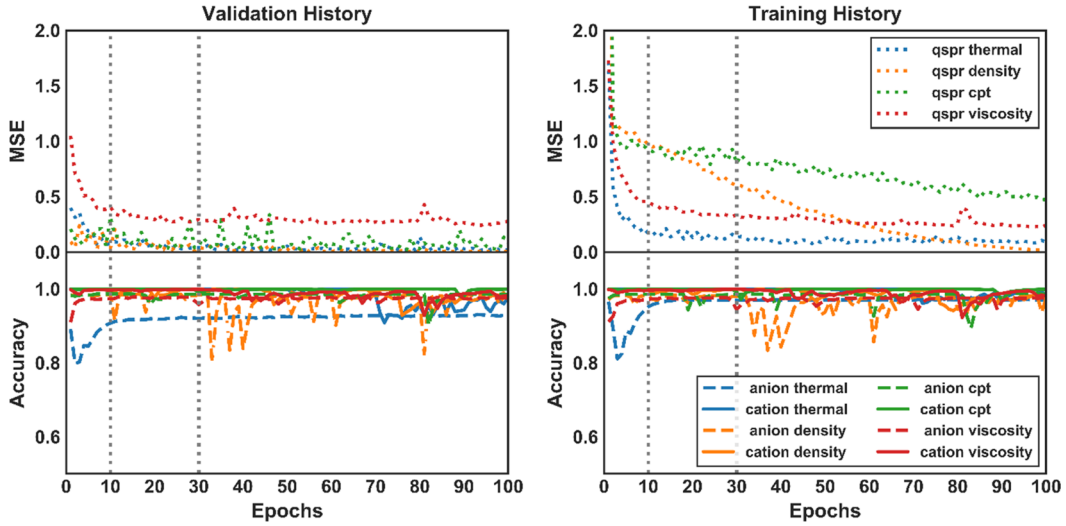


Figure 4. (left) Validation and (right) training set histories for QSPR training on Gen3. The top parts of each plot show the mean squred errors (MSEs) of the property estimations for the validation and training set, respectively, while the accuracies in the bottom half of each plot represent how accurately the ions were recreated by the networks.

alteration that the input to the QSPR consisted of the respective latent spaces.

The training histories of these three models are presented in Figure 2, right panel. The Gen3 model, with completely separate networks for the cation and anion, achieved high reconstruction accuracies in almost every epoch of training. Gen1 and Gen2 failed to improve in the first million samples of GDB-17 data (Figure 2, right panel, up to the first dotted line) and only marginally improved during the first transfer learning segment, when both GDB-17 and cation/anion data were used for training (between first and second dotted lines in Figure 2, right panel). Of note, Gen1 and Gen2 were able to achieve modestly high accuracies for anions in their final phase of training (after the second dotted line in Figure 2, right panel). This is due to the anion data set being much smaller than the cation data set (98 vs 276); i.e., the VAE was able to memorize the anion structures but was unable to generalize across molecular entities. Across all models, training accuracies on the smaller data sets in the second million samples are marked by "jumps" from the prior, larger GDB17 data sets due to the VAE's ability to memorize the structures in the training sets. The true test of the model's viability will be whether it can generate new IL structures. This is later investigated.

The purpose of exposing our models to GDB-17 data before IL data is to embed within the network rudimentary chemical understanding, insofar as to be able to recreate SMILES annotation while dealing with noise (stochastic embedding) and information loss (bottlenecking in the latent space). However, it is important for the VAEs to learn the features within various cation/anion groups when the IL data is introduced in the training, specifically during the third phase of transfer learning. Therefore, we performed principle component analysis on the cation VAE Gen3 model during the third phase of the training. Figure 3 shows the first two principle components during phase three salt embedding at every 100,000 training samples. As expected, during this phase of training, the Gen3 model assigns cation types to specific neighborhoods within the latent space. Initially, the salt components of various cationic functional groups are dispersed throughout the latent space, but as more and more samples are introduced in the network, the latent space seems to get rearranged based on type of cations. For example, there are clear regions consisting only of either the pyridinium or imidazolium type cations, for which there are ample training data. The quinolinium type cations have a distinct region as well, and the remaining cation types cluster together around the origin of the principal components. Also of interest, from the first to the fifth 100,000 training examples, the quinolinium type cations appear to migrate together in their latent embeddings. This is without exposing the VAE to any kind of "type labeling"—the VAE is learning for itself these structural categories that have been ascribed by researchers.

Next, to confirm whether these models—Gen1, Gen2, and Gen3—would be useful for generative purposes, we tasked them with generating unique structures as in the case of the cation generator, albeit this time allocating the entire salt database as seeds for the respective latent spaces and allowing the models however many function calls needed to procure 100 unique structures. Gen1 produced 100 structures in 4,073 function calls, Gen2 in 13,968, and Gen3 in 2,597. As forecasted by their poor recreation accuracies, however, the Gen1 and Gen2 models achieved low structural variety in their procured candidates. Their molecular returns were typically long, branched and unbranched alkyl chains. The Gen2 model, however, did achieve some greater structural variety than the Gen1 model: while it required far more decoding attempts (13,968 vs 4,073) to create chemically feasible structures, when it did procure a structure, it was more chemically meaningful, containing functional groups learned from its training data that were not present in the Gen1 model. The SMILES strings for these generated structures of all three generative models (Gen1, Gen2, and Gen3) are included in the Supporting Information. Due to its ability to create both heterogeneous structures and at a low function call level, Gen3 was selected in subsequent QSPR trainings.

After finalizing the training protocol (M5) and VAE network architecture (Gen3), we combined these components with QSPR training to predict various IL properties. We trained four VAE+QSPR models for four different properties of ILs: density, heat capacity, viscosity, and thermal conductivity using the Gen3 model. In this training schema, the hidden cells of Gen3, indicated in Figure 2, are unfrozen (their weights can be

https://dx.doi.org/10.1021/acs.jpcb.0c05938
J. Phys. Chem. B 2020, 124, 8347−8357

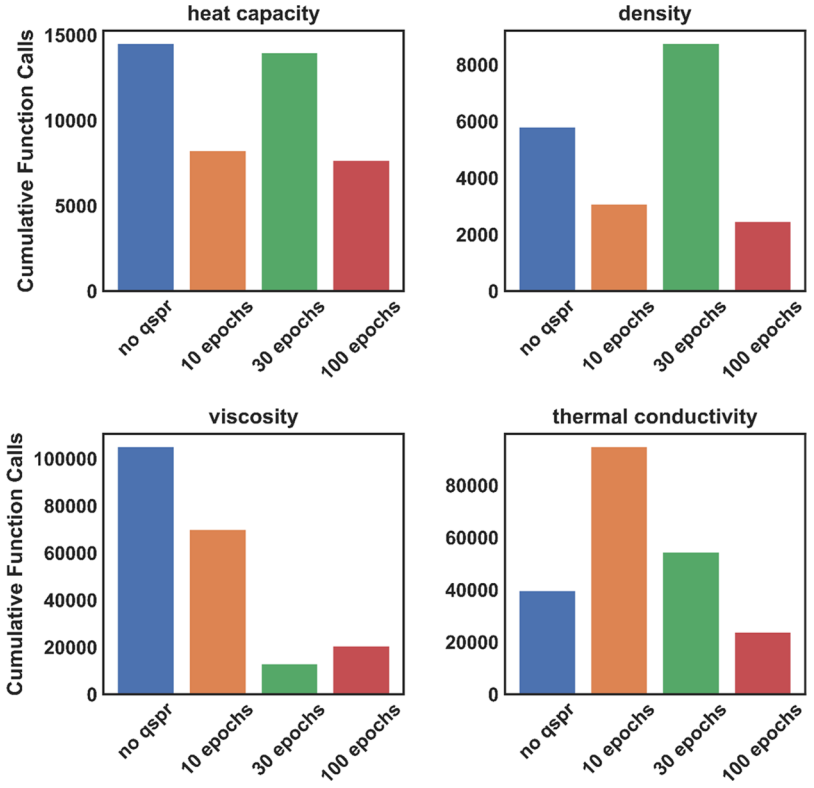


Figure 5. Cumulative function calls for each property (heat capacity, density, viscosity, and thermal conductivity) for VAE with no QSPR and the VAE-QSPR model with varying number of QSPR epochs to create 100 unique structures not within the training data sets with target property values.

updated) and the QSPR loss is added to the objective function. Instead of training on the entire cation/anion data sets, each subsequent model is dedicated to a specific property and is only trained on the subset of cation/anion data for which that property data is available. The validation and training set histories are presented in Figure 4 where MSEs represent the mean absolute error of each training, while the accuracies represent how accurately the properties were estimated by the networks. Figure 4, left panel, indicates that the Gen3 QSPR validation sets did not appreciably improve after 10 epochs and did not improve at all after 30. To avoid overfitting, a model would be selected somewhere between epoch 10 and 30, since the validation loss improves negligibly in this range. In this case, however, a model with the best generative capacity was desired, which did not necessarily correlate with its predictive ability. Indeed, the QSPR training served to redistribute the placement of chemicals embedded in the latent space, to better navigate it according to the premise that like structures lead to like properties.

To validate this generative capability and to test the usefulness of the VAE-latent space search approach, the Gen3 model was saved at 10, 30, and 100 epochs for each of the four QSPR training sessions. To simulate a real-world design criterion, we designated each of the four properties as a value to maximize or minimize. The target properties were the following: high heat capacity (>918 J/mol/K), low density (<962.7 kg/m^3), low viscosity (<0.0106 Pa s), and high thermal conductivity (>0.1667 W/m/K).

For heat capacity and thermal conductivity, the top 10 cations corresponding to the highest values for the respective properties were taken as seeds for the VAE-QSPRs. The models were then tasked with returning 100 salts with property values that were equal to or higher than the experimental salt values. The same was done for density and viscosity with the alteration that the lowest values for the respective properties were taken as seeds and models were tasked with finding salts with equal or lower values. These four VAE-QSPR tasks were repeated with the non-QSPR-trained Gen3 model to highlight whether the reorganization of the latent space improved the ability of the generator to find desired values. In order to compare the VAE-QSPRs with the original Gen3 model (VAE without QSPR), a separate RDKit-based QSPR model similar to that described in previous work[20] was used as the property evaluator. Next, the performance of each VAE with increasing training of the QSPR predictor (10, 30, and 100 epochs) was compared against the original Gen3 VAE in their ability to procure salts with property values that bordered on their respective distributions. Therefore, we ended up with 13 different generator models (three QSPR-VAEs differing in epochs for each of four properties and the original Gen3 model) all of which were tasked with generating 100 salt structures with the aforementioned target properties. The total number of function calls required by each generator was recorded. The total number of function calls for each QSPR-VAE and the baseline Gen3 VAE are presented in Figure 5.

In Figure 5, we observe that the VAE with 100 epochs, on average, performs the best out of the VAE models. This suggests that, by grafting a QSPR predictor onto an existing generative model, the latent space—previously organized by a purely structural relationship—is reorganized by the goal of minimizing QSPR loss. By using the top 10 ILs as seeds to generate from the latent space, we leverage the QSPR-related organization but without explicitly calculating a gradient. In the next portion of this work, we investigate how calculating the

https://dx.doi.org/10.1021/acs.jpcb.0c05938
J. Phys. Chem. B 2020, 124, 8347−8357

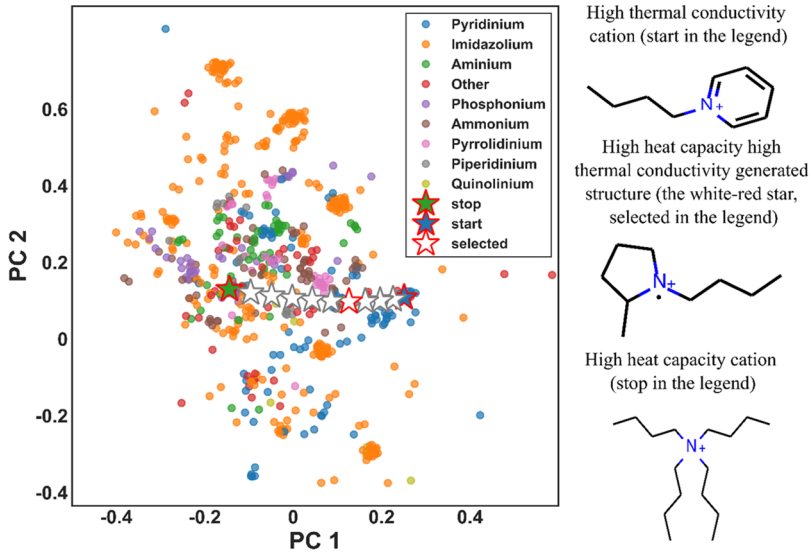


Figure 6. Latent space interpolation in the Gen3 thermal conductivity−heat capacity 100 epoch QSPR model. In the interpolation (represented by stars), the blue-red star indicates a high thermal conductivity salt in the training data and the green-red star indicates a high heat capacity in the training data. The white stars represent the interpolated structures, and the white-red star represents the selected structure with both high heat capacity and high thermal conductivity

spherical linear interpolation (SLERP) between ILs embedded in the latent space can lead to new ILs with combinative macroscopic properties.

Interpolating in the Latent Space for Combinative Properties. A distinct advantage of designing in a continuous structural space is the arbitrary appropriation of mathematical relationships between vector-represented molecules to procure new ones. However, this interpolation requires calculating distances in high dimensional feature spaces, which is nontrivial. Often, we will have to actively avoid pockets of data-scarce regions in the VAE. Specifically, sampling from a spherical linear interpolation (SLERP)[54] between embeddings rather than a linear one helps prevent divergence from a model's prior distribution (the distribution in the latent space).[62] Indeed, SLERP is just one of a handful of sampling techniques used in high dimensional latent space models.[62] In other areas of generative modeling research, SLERP has been used to demonstrate that the model has not simply memorized training data but can extrapolate outside known examples.[33]

SLERP is a method to interpolate between two vectors along the shortest arc.[54] It can be thought of as the shortest path along a spherical geodesic. More specifically, in the context of unit vectors (which we can extend to any vectors by normalizing), it is the interpolation between two unit vectors along a unit-radius great circle arc centered at the origin, with constant-speed (angular velocity) motion. Originally developed for the purposes of quaternion interpolation for 3D animation, SLERP can be defined and used independently of quaternions and beyond their dimensionality (4D vectors along a 4D hypersphere). In the context of this paper, we interpolate between n-dimensional vectors along an n-dimensional sphere, where n is the dimensionality of the latent space.

The formula for SLERP interpolation between two vectors p_0 and p_1(normalized) is independent of the dimensionality of the space in which the arc is subtended and depends on an interpolation parameter t between 0 and 1, as well as Ω, the angle subtended between the points such that $\cos(\Omega) = p_0 \cdot p_1$, the n-dimensional dot product:

$$\mathrm{SLERP}(p_0, p_1; t) = \frac{\sin[(1-t)\Omega]}{\sin(\Omega)} p_0 + \frac{\sin[t\Omega]}{\sin(\Omega)} p_1$$

The interpolation t dictates the point on the arc to which the interpolation is set. For example, $t = 0.10$ refers to a point 10% of the way from p_0 to p_1, such that the angle between the interpolated point and p_0 is 0.1Ω and the angle between the interpolated point and p_1 is 0.9Ω. Changing t at a constant rate results in a constant angular velocity along the arc. One will also notice that, in the limit of $\Omega \to 0$, the formula or SLERP reduces to that of linear interpolation (LERP):

$$\mathrm{LERP}(p_0, p_1; t) = (1-t)p_0 + tp_1$$

In this phase of our schema, we trained six combinative property models containing any two of the four thermophysical properties. These models were trained in the same way as before, albeit with two property targets for the QSPR layers trained off the latent space. After training, to evaluate the performance of these models, we would at any given iteration select two cationic moieties from the top 20 pool of each property and interpolate 10 structures between them using SLERP coordinates. The model was allowed 100 sampling attempts before moving on past the current interpolation (i.e., in an iteration call, 0−10 structures were returned). Figure 6 demonstrates one such example when Gen3 VAE is used to interpolate between two embedded ILs with desirable properties, namely, high heat capacity and thermal conductivity, which is often required for heat transfer application. In this search, a high thermal conductivity IL was selected with a high heat capacity IL, which appears at the top and bottom of the right panel in Figure 6. After interpolating between them for 10 distinct structures, a candidate was found that had heat capacity and thermal conductivity estimates within our dictated cutoff (higher than 918 J/mol/K and 0.1667 W/m/K). This structure is shown as the white-red star in Figure 6. Similar molecules to our selected candidate, essentially a pyrrole radical, have been synthesized.[63] The radical presence has been attributed to the pyrrole ring, in which it acts as an effective free-radical trap.[63,64] Aminium or six-membered ring

https://dx.doi.org/10.1021/acs.jpcb.0c05938
J. Phys. Chem. B 2020, 124, 8347−8357

radical ions have been reported as reaction pathways for creation of salts under certain conditions.[64]

To quantify whether interpolation is a convenient search mechanism, we tallied the total number of function calls to procure 10 structures in each of the pairwise property profiles, as shown in Table 3. After procuring the interpolated structures, they were evaluated alongside a sampled anion (without anion pairing in Table 3) or alongside all experimental anions (with anion pairing in Table 3). In this way, total VAE function calls were minimized. If the candidate−anion pair was estimated to be within the property target bounds, it was selected as a solution to the search criteria. In addition, we performed the sampling without interpolation, instead using a latent "noisy" seed. The results of these sample attempts appear in the right two columns of Table 3.

Table 3. Total Function Calls to Procure 10 Candidate IL Materials within the Specified Property Targets

	interpolative function calls		noisy seed function calls	
properties	without anion pairing	with anion pairing	without anion pairing	with anion pairing
heat capacity, thermal conductivity	13198	234	1438	340
heat capacity, viscosity	86288	5817	38084	1006
heat capacity, density	11727	467	604	27
density, thermal conductivity	2037	712	1161	208
density, viscosity	161118	15900	66597	2263
viscosity, thermal conductivity	14920	1599	19334	2260

As seen by comparing the right and left sides of Table 3, sampling from the top 20 performers (encoded as latent seeds) for each property outperformed the interpolation strategy. However, when procured candidates were evaluated against all anions in the training data sets to evaluate valid anionic partners for the given property distribution target, the interpolation method performed better than the noisy seed strategy for generating high heat capacity/high thermal conductivity targets and low viscosity/high thermal conductivity targets. The reason evaluating against all anion partners results in lower function calls is that the cations are often promiscuous—they can be attached to a number of anions and still fit within the target property profile.

CONCLUSION

In this work, we have demonstrated the following objectives: (1) transfer learning is an effective approach to create a generative neural network model of molecule types for which there is scarce training data and this approach works for generalizing from single to dual molecular systems (i.e., ILs), (2) training the preconditioned-structural model on subsequent property data can lead to effective reorganization of the latent space for generating molecules with desired properties, and (3) interpolating between molecules of property extremes can result in hybrid-generated structures with structural and property similarities to the two end points. For the first objective, we evaluated our training protocol against null hypotheses: (a) training a model on only the larger but molecularly dissimilar data set and (b) training the model on only the smaller, target data set. Both null hypothesis protocols resulted in poor structural variety when sampling from the generated model and/or minimal chemically feasible SMILES. In the second objective, we tasked our generative models with producing 100 chemically feasible structures at fringe property distributions (i.e., a nontrivial design task) and found that QSPR-trained models procured structural candidates with fewer function calls than the non-QSPR trained model. Finally, for the third objective, we outlined a multiobjective design problem, where we sought fringe property distributions using six pairwise property combinations. In these cases, interpolating between top performers from each distribution did produce chemically feasible structures but at about the same performance as sampling from both distributions with noisy seeds.

The infeasibility of purely MD/MC approaches to exploring small molecule space in search of IL candidates was stated earlier. A future area of work, however, would be to screen the small molecule candidates procured by the ML approach, with quantum mechanical structure relaxation followed by MD/MC simulations to determine the viability of the structures along with their properties. Additionally, one of the main challenges of this study was the availability of cation and anion data in the ILThermo database. Though our transfer learning protocol showed significant promise even with a very small amount of data for a relatively new domain such as ionic liquids, there are other paths one could take to build a reasonably large data set of anions and cations. These paths include but are not limited to selecting representative molecules either from the Zinc[46] or PubChem[65] database and generating rule-based libraries of possible cations/anions such as replacing an aromatic group of small aromatic cationic rings with an alkyl group. This will help the transfer learning protocol proposed in this study to learn more complex relationships in the latent space, as the network will observe many more samples and may generate a more diverse set of molecules. Additionally, comparison of the quality and size of rule-based libraries with the VAE approach discussed in this work and other machine learning approaches would help contextualize the utility of machine learning approaches, generally.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.0c05938.

Figures S1−S4 containing samples of positive structures produced by different models; tabulation of the SMILES strings of 100 unique salt structures and the number of function calls (attempts) to generate them for each of the generative models (Gen1, Gen2, and Gen3) (Table S1) (PDF)

AUTHOR INFORMATION

Corresponding Author

Jim Pfaendtner − *Department of Chemical Engineering, University of Washington, Seattle, Washington 98105, United States;* orcid.org/0000-0001-6727-2957; Email: jpfaendt@uw.edu

https://dx.doi.org/10.1021/acs.jpcb.0c05938
J. Phys. Chem. B 2020, 124, 8347−8357

Authors

Wesley Beckner − *Department of Chemical Engineering, University of Washington, Seattle, Washington 98105, United States*

Chowdhury Ashraf − *Department of Chemical Engineering, University of Washington, Seattle, Washington 98105, United States*

James Lee − *Department of Chemical Engineering, University of Washington, Seattle, Washington 98105, United States*

David A. C. Beck − *Department of Chemical Engineering, University of Washington, Seattle, Washington 98105, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.0c05938

Author Contributions

†W.B. and C.A. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

W.B. and J.P. would like to acknowledge the funding from NRT-DESE: Data Intensive Research Enabling Clean Technologies (DIRECT) under Grant No. NSF #1633216. C.A., D.A.C.B., and J.P. would like to acknowledge support from HDR: I-DIRSE-FW: Accelerating the Engineering Design and Manufacturing Life-Cycle with Data Science under Grant No. NSF #1934292. This work was facilitated using computational, storage, and networking infrastructure provided by the Hyak supercomputer system, supported by the University of Washington.

■ REFERENCES

(1) Kar, M.; Tutusaus, O.; MacFarlane, D. R.; Mohtadi, R. Novel and Versatile Room Temperature Ionic Liquids for Energy Storage. *Energy Environ. Sci.* **2019**, *12* (2), 566−571.

(2) Plechkova, N. V.; Seddon, K. R. Applications of Ionic Liquids in the Chemical Industry. *Chem. Soc. Rev.* **2008**, *37*, 123−150.

(3) Xiao, M.; Liu, H.; Gao, H.; Olson, W.; Liang, Z. CO2 Capture with Hybrid Absorbents of Low Viscosity Imidazolium-Based Ionic Liquids and Amine. *Appl. Energy* **2019**, *235*, 311−319.

(4) Ding, S.; Guo, Y.; Hülsey, M. J.; Zhang, B.; Asakura, H.; Liu, L.; Han, Y.; Gao, M.; Hasegawa, J. ya; Qiao, B.; et al. Electrostatic Stabilization of Single-Atom Catalysts by Ionic Liquids. *Chem.* **2019**, *5* (12), 3207−3219.

(5) Mena, I. F.; Diaz, E.; Pérez-Farías, C.; Stolte, S.; Moreno-Andrade, I.; Rodriguez, J. J.; Mohedano, A. F. Catalytic Wet Peroxide Oxidation of Imidazolium-Based Ionic Liquids: Catalyst Stability and Biodegradability Enhancement. *Chem. Eng. J.* **2019**, *376*, 120431.

(6) Kasprzak, D.; Krystkowiak, E.; Stępniak, I.; Galiński, M. Dissolution of Cellulose in Novel Carboxylate-Based Ionic Liquids and Dimethyl Sulfoxide Mixed Solvents. *Eur. Polym. J.* **2019**, *113*, 89−97.

(7) Kaur, N. Ionic Liquids: A Versatile Medium for the Synthesis of Six-Membered Two Nitrogen- Containing Heterocycles. *Curr. Org. Chem.* **2019**, *23* (1), 76−96.

(8) Tian, Y. H.; Goff, G. S.; Runde, W. H.; Batista, E. R. Exploring Electrochemical Windows of Room-Temperature Ionic Liquids: A Computational Study. *J. Phys. Chem. B* **2012**, *116* (39), 11943−11952.

(9) Di Pietro, M. E.; Margola, T.; Celebre, G.; De Luca, G.; Saielli, G. A Combined LX-NMR and Molecular Dynamics Investigation of the Bulk and Local Structure of Ionic Liquid Crystals. *Soft Matter* **2019**, *15* (22), 4486−4497.

(10) Sprenger, K. G.; Jaeger, V. W.; Pfaendtner, J. The General AMBER Force Field (GAFF) Can Accurately Predict Thermodynamic and Transport Properties of Many Ionic Liquids. *J. Phys. Chem. B* **2015**, *119* (18), 5882−5895.

(11) Beckner, W.; Pfaendtner, J. Fantastic Liquids and Where to Find Them: Optimizations of Discrete Chemical Space. *J. Chem. Inf. Model.* **2019**, *59* (6), 2617−2625.

(12) Hosseinzadeh, M.; Hemmati-Sarapardeh, A. Toward a Predictive Model for Estimating Viscosity of Ternary Mixtures Containing Ionic Liquids. *J. Mol. Liq.* **2014**, *200* (PB), 340−348.

(13) Venkatraman, V.; Evjen, S.; Knuutila, H. K.; Fiksdahl, A.; Alsberg, B. K. Predicting Ionic Liquid Melting Points Using Machine Learning. *J. Mol. Liq.* **2018**, *264*, 318−326.

(14) Hezave, A. Z.; Raeissi, S.; Lashkarbolooki, M. Estimation of Thermal Conductivity of Ionic Liquids Using a Perceptron Neural Network. *Ind. Eng. Chem. Res.* **2012**, *51* (29), 9886−9893.

(15) Paduszyński, K.; Domańska, U. Viscosity of Ionic Liquids: An Extensive Database and a New Group Contribution Model Based on a Feed-Forward Artificial Neural Network. *J. Chem. Inf. Model.* **2014**, *54* (5), 1311−1324.

(16) Torrecilla, J. S.; Rodríguez, F.; Bravo, J. L.; Rothenberg, G.; Seddon, K. R.; López-Martin, I. Optimising an Artificial Neural Network for Predicting the Melting Point of Ionic Liquids. *Phys. Chem. Chem. Phys.* **2008**, *10* (38), 5826−5831.

(17) Kruglov, I.; Sergeev, O.; Yanilkin, A.; Oganov, A. R. Energy-Free Machine Learning Force Field for Aluminum. *Sci. Rep.* **2017**, *7* (1), 1−7.

(18) Botu, V.; Batra, R.; Chapman, J.; Ramprasad, R. Machine Learning Force Fields: Construction, Validation, and Outlook. *J. Phys. Chem. C* **2017**, *121* (1), 511−522.

(19) Huan, T. D.; Batra, R.; Chapman, J.; Krishnan, S.; Chen, L.; Ramprasad, R. A Universal Strategy for the Creation of Machine Learning-Based Atomistic Force Fields. *Npj Comput. Mater.* **2017**, *3* (1), 37.

(20) Beckner, W.; Mao, C. M.; Pfaendtner, J. Statistical Models Are Able to Predict Ionic Liquid Viscosity across a Wide Range of Chemical Functionalities and Experimental Conditions. *Mol. Syst. Des. Eng.* **2018**, *3* (1), 253−263.

(21) Elton, D. C.; Boukouvalas, Z.; Fuge, D.; Chung, P. W. Deep Learning for Molecular Design—a Review of the State of the Art. *Mol. Syst. Des. Eng.* **2019**, *4*, 828.

(22) Dahl, G. E.; Jaitly, N.; Salakhutdinov, R. Multi-Task Neural Networks for QSAR Predictions. 2014, arXiv:1406.1231. arXiv.org e-Print archive. https://arxiv.org/abs/1406.1231.

(23) NVIDIA DGX-1 With Tesla V100 System Architecture The Fastest Platform for Deep Learning. https://images.nvidia.com/content/pdf/dgx1-v100-system-architecture-whitepaper.pdf (accessed Jun 24, 2020).

(24) Gulli, A.; Pal, S. *Deep Learning with Keras*; Packt Publishing Ltd: Birmingham, U.K., 2017.

(25) Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M. Tensor Flow: A System for Large-Scale Machine Learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016*; 2016; 265−283.

(26) Lake, B. M.; Salakhutdinov, R.; Tenenbaum, J. B. Human-Level Concept Learning through Probabilistic Program Induction. *Science* **2015**, *350* (6266), 1332−1338.

(27) Belhaj, M.; Protopapas, P.; Pan, W. Deep Variational Transfer: Transfer Learning through Semi-Supervised Deep Generative Models. 2018, arXiv:1812.03123. arXiv.org e-Print archive. https://arxiv.org/abs/1812.03123.

(28) Zhang, R.; Isola, P.; Efros, A. A. Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2016; pp 1058−1067.

(29) Goh, G. B.; Vishnu, A.; Siegel, C.; Hodas, N. Using Rule-Based Labels for Weak Supervised Learning: A ChemNet for Transferable Chemical Property Prediction. In *Proceedings of the ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining; ACM: 2018; pp 302−310. DOI: 10.1145/3219819.3219838.
(30) Goh, G. B.; Hodas, N. O.; Siegel, C.; Vishnu, A. SMILES2Vec: An Interpretable General-Purpose Deep Neural Network for Predicting Chemical Properties. **2017**, arXiv:1712.02034. arXiv.org e-Print archive. https://arxiv.org/abs/1712.02034.
(31) Fare, C.; Turcani, L.; Pyzer-Knapp, E. O. Powerful, Transferable Representations for Molecules through Intelligent Task Selection in Deep Multitask Networks. *Phys. Chem. Chem. Phys.* **2020**, *22*, 13041.
(32) Donahue, J.; Hendricks, L. A.; Rohrbach, M.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; Darrell, T. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39* (4), 677−691.
(33) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268−276.
(34) Goh, G. B.; Vishnu, A.; Siegel, C.; Hodas, N. Using Rule-Based Labels for Weak Supervised Learning: A ChemNet for Transferable Chemical Property Prediction. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: 2018; pp 302−310. DOI: 10.1145/3219819.3219838.
(35) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4* (1), 120−131.
(36) Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J. L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52* (11), 2864−2875.
(37) Zhao, S.; Song, J.; Ermon, S. Towards Deeper Understanding of Variational Autoencoding Models. **2017**, arXiv:1702.08658. arXiv.org e-Print archive. https://arxiv.org/abs/1702.08658.
(38) Larsen, A. B. L.; Sønderby, S. K.; Larochelle, H.; Winther, O. Autoencoding beyond Pixels Using a Learned Similarity Metric. In *33rd International Conference on Machine Learning, ICML 2016*; 2016; Vol. *4*, pp 2341−2349.
(39) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*; 2014.
(40) Popova, M.; Shvets, M.; Oliva, J.; Isayev, O. Molecular RNN: Generating Realistic Molecular Graphs with Optimized Properties. **2019**, arXiv:1905.13372. arXiv.org e-Print archive. https://arxiv.org/abs/1905.13372.
(41) Jin, W.; Barzilay, R.; Jaakkola, T. Hierarchical Graph-to-Graph Translation for Molecules. **2019**, arXiv:1907.11223. arXiv.org e-Print archive. https://arxiv.org/abs/1907.11223.
(42) Winter, R.; Montanari, F.; Noé, F.; Clevert, D. A. Learning Continuous and Data-Driven Molecular Descriptors by Translating Equivalent Chemical Representations. *Chem. Sci.* **2019**, *10* (6), 1692−1701.
(43) Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31−36.
(44) O'Boyle, N. M. Towards a Universal SMILES Representation - A Standard Method to Generate Canonical SMILES Based on the InChI. *J. Cheminf.* **2012**, *4* (9), 22.
(45) Koichi, S.; Iwata, S.; Uno, T.; Koshino, H.; Satoh, H. Algorithm for Advanced Canonical Coding of Planar Chemical Structures That Considers Stereochemical and Symmetric Information. *J. Chem. Inf. Model.* **2007**, *47* (5), 1734−1746.
(46) Virshup, A. M.; Contreras-García, J.; Wipf, P.; Yang, W.; Beratan, D. N. Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-like Compounds. *J. Am. Chem. Soc.* **2013**, *135* (19), 7296−7303.
(47) Gaulton, A.; Hersey, A.; Patr, A.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibri, E.; Davies, M.; et al. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2016**, *45* (D1), D945−D954.
(48) Noh, J.; Kim, J.; Stein, H. S.; Sanchez-Lengeling, B.; Gregoire, J. M.; Aspuru-Guzik, A.; Jung, Y. Inverse Design of Solid-State Materials via a Continuous Representation. *Matter* **2019**, *1* (5), 1370−1384.
(49) Winter, R.; Montanari, F.; Noé, F.; Clevert, D. A. Learning Continuous and Data-Driven Molecular Descriptors by Translating Equivalent Chemical Representations. *Chem. Sci.* **2019**, *10* (6), 1692−1701.
(50) Samanta, B.; De, A.; Jana, G.; Chattaraj, P. K.; Ganguly, N.; Rodriguez, M. G. NeVAE: A Deep Generative Model for Molecular Graphs. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33* (01), 1110−1117.
(51) Landrum, G. RDKit: Open-source cheminformatics. http://www.rdkit.org/.
(52) Gregor, K.; Danihelka, I.; Graves, A.; Rezende, D. J.; Wierstra, D. DRAW: A Recurrent Neural Network for Image Generation. In *32nd International Conference on Machine Learning, ICML 2015*; 2015; Vol. *2*, pp 1462−1471.
(53) Doersch, C. Tutorial on Variational Autoencoders. **2016**, arXiv:1606.05908. arXiv.org e-Print archive. https://arxiv.org/abs/1606.05908.
(54) Shoemake, K. Animating Rotation with Quaternion Curves. *Comput. Graph.* **1985**, *19* (3), 245−254.
(55) Kazakov, A.; Magee, J. W.; Chirico, R. D.; Paulechka, E.; Diky, V.; Muzny, C. D.; Kroenlein, K.; Frenkel, M. NIST Standard Reference Database 147: NIST Ionic Liquids Database-(ILThermo). 2013.
(56) Dong, Q.; Muzny, C. D.; Kazakov, A.; Diky, V.; Magee, J. W.; Widegren, J. A.; Chirico, R. D.; Marsh, K. N.; Frenkel, M. ILThermo: A Free-Access Web Database for Thermodynamic Properties of Ionic Liquids. *J. Chem. Eng. Data* **2007**, *52*, 1151−1159.
(57) Belayneh, A.; Adamowski, J.; Khalil, B.; Quilty, J. Coupling Machine Learning Methods with Wavelet Transforms and the Bootstrap and Boosting Ensemble Approaches for Drought Prediction. *Atmos. Res.* **2016**, *172*−*173*, 37−47.
(58) Erdal, H. I.; Karakurt, O. Advancing Monthly Streamflow Prediction Accuracy of CART Models Using Ensemble Learning Paradigms. *J. Hydrol.* **2013**, *477*, 119−128.
(59) Tiwari, M. K.; Chatterjee, C. A New Wavelet-Bootstrap-ANN Hybrid Model for Daily Discharge Forecasting. *J. Hydroinf.* **2011**, *13* (3), 500−519.
(60) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273−1280.
(61) Sakloth, K.; Beckner, W.; Pfaendtner, J.; Goh, G. B. IL-Net: Using Expert Knowledge to Guide the Design of Furcated Neural Networks. In *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*; IEEE: 2019; pp 1465−1473. DOI: 10.1109/BigData.2018.8622512.
(62) White, T. Sampling Generative Networks. **2016**, arXiv:1609.04468. arXiv.org e-Print archive. https://arxiv.org/abs/1609.04468.
(63) Gritter, R. J.; Chriss, R. J. Free-Radical Reactions of Pyrroles. *J. Org. Chem.* **1964**, *29* (5), 1163−1167.
(64) Boursalian, G. B.; Ham, W. S.; Mazzotti, A. R.; Ritter, T. Charge-Transfer-Directed Radical Substitution Enables Para-Selective C-H Functionalization. *Nat. Chem.* **2016**, *8* (8), 810−815.
(65) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47*, D1102.

https://dx.doi.org/10.1021/acs.jpcb.0c05938
J. Phys. Chem. B 2020, 124, 8347−8357