International Journal of
Molecular Sciences

MDPI

Article

Deep Probabilistic Learning Model for Prediction of Ionic Liquids Toxicity

Mapopa Chipofya [1], Hilal Tayara [2,*] and Kil To Chong [1,3,*]

[1] Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju 54896, Korea; mapopachipofya@jbnu.ac.kr
[2] School of International Engineering and Science, Jeonbuk National University, Jeonju 54896, Korea
[3] Advanced Electronics and Information Research Center, Jeonbuk National University, Jeonju 54896, Korea
* Correspondence: hilaltayara@jbnu.ac.kr (H.T.); kitchong@jbnu.ac.kr (K.T.C.); Tel.: +82-63-270-2478 (K.T.C.)

Abstract: Identification of ionic liquids with low toxicity is paramount for applications in various domains. Traditional approaches used for determining the toxicity of ionic liquids are often expensive, and can be labor intensive and time consuming. In order to mitigate these limitations, researchers have resorted to using computational models. This work presents a probabilistic model built from deep kernel learning with the aim of predicting the toxicity of ionic liquids in the leukemia rat cell line (IPC-81). Only open source tools, namely, RDKit and Mol2vec, are required to generate predictors for this model; as such, its predictions are solely based on chemical structure of the ionic liquids and no manual extraction of features is needed. The model recorded an RMSE of 0.228 and R^2 of 0.943. These results indicate that the model is both reliable and accurate. Furthermore, this model provides an accompanying uncertainty level for every prediction it makes. This is important because discrepancies in experimental measurements that generated the dataset used herein are inevitable, and ought to be modeled. A user-friendly web server was developed as well, enabling researchers and practitioners ti make predictions using this model.

Keywords: small molecules; ionic liquids; toxicity; probabilistic deep learning; artificial intelligence

Citation: Chipofya, M.; Tayara, H.; Chong, K.T. Deep Probabilistic Learning Model for Prediction of Ionic Liquids Toxicity. *Int. J. Mol. Sci.* **2022**, *23*, 5258. https://doi.org/10.3390/ijms23095258

Academic Editors: Johannes Kirchmair and Ya Chen

Received: 3 April 2022
Accepted: 6 May 2022
Published: 9 May 2022

1. Introduction

Materials which exist in liquid phase at temperatures below 100 °C and are composed of organic or inorganic cations and anions are referred to as room temperature ionic liquids. Often, they are more loosely called ionic liquids (ILs). These materials exhibit a unique set of desirable properties, such as a low melting point, negligible volatility, thermal and chemical stability, high ionic conductivity, solubility with many compounds, low flammability, moderate viscosity, high polarity, and high recyclability [1–4]. Hence, they have drawn great interest as a research topic and found applications in various fields such as catalysis [5,6], pharmaceuticals [7,8], biopolymer processing [9], nuclear fuel reprocessing [10,11], solar thermal energy [12], and batteries [2,13]. However, there is a concern that, owing to their solubility in aqueous media, ionic liquids may interact with biota, distress it, and ultimately impact human health when these chemicals are discharged into the environment through wastewater [14].

Prominent research results regarding the toxic effects induced by ILs in the ecosystem are presented in the works of Samorì et al. [15] and Latała et al. [16]. Overall, studies leading to identification of more ILs with known effects on the environment have increased at a slower pace than anticipated [14]. The usual and most effective way of conducting experiments to measure the toxicity of ILs directly with the aim of determining ILs with desirable low toxicity has been deemed time-consuming, resource-intensive, and even impractical due to the large number of feasible combinations between cations and anions [14,17]. To quickly build on the available results obtained from experimental measurements and mitigate the limitations associated with conducting further experimental

measurements, computational methods, which often involve machine learning, have become a preferred tool. Herein, we consider several computational tools that have been developed recently to predict the toxicity of ILs against the leukemia rat cell line IPC-81. IPC-81 has been frequently used to quantitatively indicate the toxicity of ILs [14,18–25].

Wang et al. [17] developed a support vector machine (SVM) model based on a dataset containing 355 ILs. From their respective simplified molecular-input line entry system (SMILES) strings, nine cation descriptors, nine anion descriptors, and 24 general descriptors were obtained for each IL using a feature extraction algorithm [26] and the RDKit cheminformatics tool [27]. Their feature extraction algorithm uses a predefined set of substructures which act as descriptors. The frequency with which each descriptor appears in the IL molecule is then used as input to the model, similar to group contribution (GC)-based methods [28–32]. The SVM model trained in this way yielded a satisfactory RMSE of 0.2875 on the 355 ILs.

More recently, Kang et al. [33] embarked on improving traditional GC-based approaches to predicting the toxicity of ionic liquids [34,35]. They developed a novel method, termed atom surface fragment contribution (ASFC), which uses the surface area of screening charge density ($S_{\sigma\text{-surface}}$) calculated based on quantum chemistry. Unlike in GC, where only the types and frequencies of functional groups are considered and interactions between groups are ignored (thus rendering isomeric groups indistinguishable [34]), ASFC has the capability to distinguish the contributions of each group in different molecules, and hence the potential to improve the reliability of GC models [33]. In ASFC, the $S_{\sigma\text{-surface}}$ values of atoms are obtained using BIOVIA COSMOtherm 2020 software, which contains COSMO files of 74 cations and 15 anions from the quantum chemical level of BP-TZVPD-FINE. The $S_{\sigma\text{-surface}}$ values for groups were found by summing the $S_{\sigma\text{-surface}}$ of all atoms in each group. Group $S_{\sigma\text{-surface}}$ values were used as predictor descriptors in a multiple linear regression (MLR) model similar to the one used by Hossain et al. [36]. The R^2 and MSE of the ASFC model were 0.924 and 0.071, respectively.

The models described above have shown an exceptional ability to predict toxicity with great accuracy and reliability by taking into account expert information regarding the creation of predictor descriptors. However, it may be difficult for someone who has no or little domain expertise to create such specialized descriptors in order to use them when making predictions concerning new ILs. Second, several of the models described above employed commercial software such as COSMOtherm to extract the desired features, which adds to their cost. Lastly, all these models are deterministic; they do not model uncertainty in either the data nor in the models themselves. Kang et al. [33] noted that there might be experimental errors in the set of ionic liquids that they used in their work. It is therefore crucial that the uncertainty associated with the data be included in the model.

Consequently, this work aims to achieve three main goals. First, we intend to use existing open-source software to generate descriptors for predicting toxicity of ionic liquids towards the leukemia rat cell line in a way that requires no or very little domain expertise. Second, based on these features, we intend to build an accurate and reliable probabilistic deep learning model for predicting toxicity. Such a model should be capable of capturing aleatoric uncertainty, which is the uncertainty due to irreducible noise in the data. Aleatoric uncertainty models the stochastic nature of the process of generating data [37]. Lastly, we built a web tool for the ensuing model to allow other researchers and practitioners to use it in their work.

2. Materials and Methods

2.1. Data Preparation

A dataset containing 155 ionic liquids which exhibit toxicity towards the leukemia rat cell line IPC-81 was collected from the literature [33,38]. The logarithm of half maximal effective concentration, $\log EC_{50}$, was used to represent the toxicity level, whereas the SMILES string for each ionic liquid was used to generate the features used for modeling.

The dataset was split randomly into subsets, which contained 140 ionic liquids for training and cross-validation and 15 for testing. Figure 1 depicts the overall process.

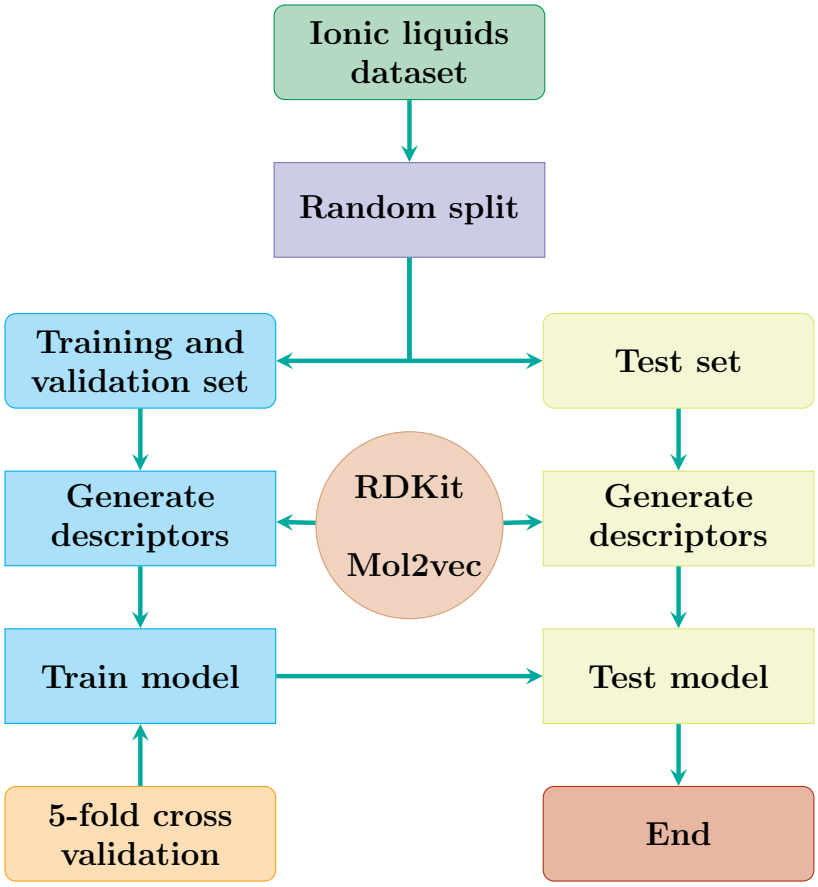

Figure 1. Workflow showing how the data were split, features generated, and cross validation used in modeling toxicity prediction for ionic liquids.

2.2. Molecular Descriptors and Features

A total of 310 features were used to describe the physical and chemical properties of each of the ionic liquids. In particular, the first ten features were obtained from RDKit molecular descriptors. These descriptors were the number of atoms in the molecule, number of heavy atoms, number of carbon atoms, number of oxygen atoms, number of nitrogen atoms, number of chlorine atoms, the topological polar surface area (TPSA) of the molecule, the molecular weight, the number of valence electrons, and the number of heteroatoms for a molecule. The rest of the features (300) were obtained using a pretrained Mol2vec [39] model. Mol2vec is an unsupervised machine learning approach to learning the vector representations of molecular substructures. The pretrained Mol2vec model used in this experiment was reported to have been trained in an unsupervised fashion on 19.9 million compounds from the ZINC version 15 [40,41] and chEMBL version 23 [42] databases. The ten features from RDkit and the 300 features from Mol2vec were then concatenated to produce one feature vector with a length of 310. Figure 2 depicts the workflow for generating these 310 features for each ionic liquid.

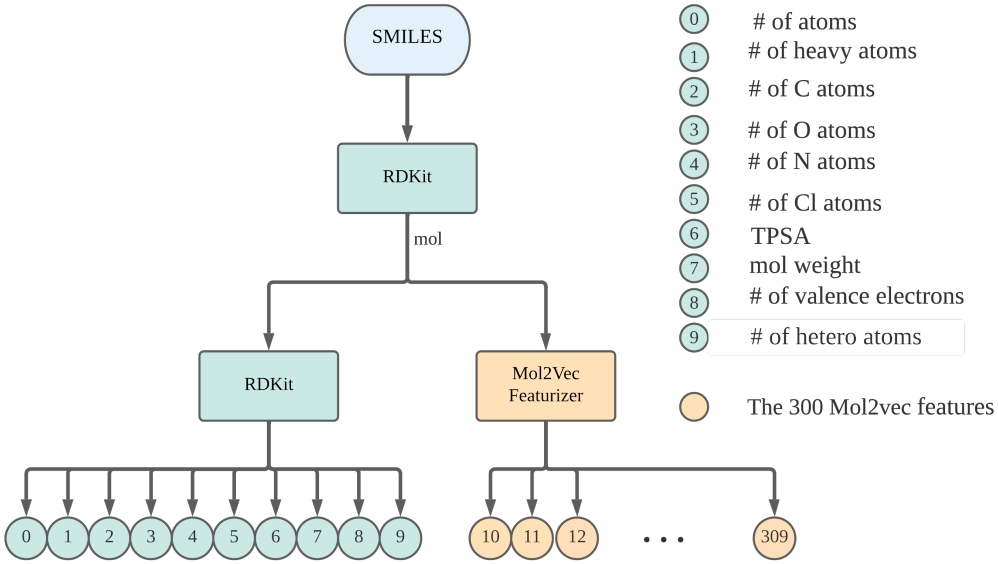

Figure 2. Workflow showing how ten features were generated from RDKit and 300 from Mol2vec for each of the ionic liquids based on their respective SMILES strings.

2.3. Deep Kernel Learning

A deep kernel model can be thought of as applying a Gaussian process with a base kernel $k\boldsymbol{\theta}$ to the final hidden layer of the deep neural network. In effect, this means that the deep neural network has a hidden layer with an infinite number of hidden units, as a Gaussian process with a base kernel $k\boldsymbol{\theta}$, such as the radial basis function (RBF) kernel, corresponds to an infinite basis function representation [43]. Figure 3 shows the pedagogical architecture of the deep kernel learning model used in our experiments.

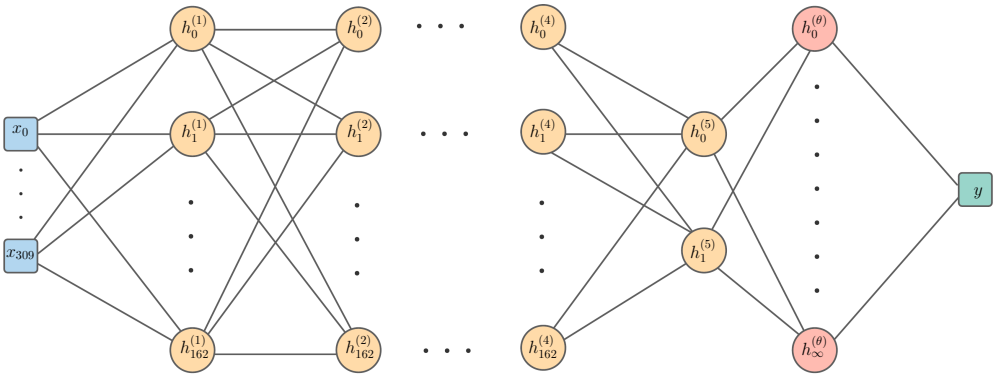

Figure 3. A Gaussian process with a deep kernel which maps the 310 input features $\mathbf{x}$ through five parametric hidden layers followed by a single hidden layer with an infinite number of basis functions using the RBF base kernel. The kernel's hyperparameters are denoted as $\boldsymbol{\theta}$, whereas those of the parametric layers are denoted as $\mathbf{w}$. Each of the first four parametric hidden layers has 163 units, while the final parametric hidden layer has two units. There is only one unit in the output $\mathbf{y}$, representing a single value for $\mathbf{log}\,\mathrm{EC}_{\mathbf{50}}$.

From an RBF base kernel $k(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta})$ with parameters $\boldsymbol{\theta}$, the input features $\mathbf{x}$ are transformed, using a probabilistic Gaussian process, as

$$k(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta}) \rightarrow k(g(\mathbf{x}_i, \mathbf{w}), g(\mathbf{x}_j, \mathbf{w})|\boldsymbol{\theta}, \mathbf{w}) \quad (1)$$

where $g(\mathbf{x}, \mathbf{w})$ is the nonlinear mapping provided by the deep neural network. The hyperparameters of the deep neural network, $\mathbf{w}$, and those of the base kernel, $\boldsymbol{\theta}$, are combined as $\gamma = \{\mathbf{w}, \boldsymbol{\theta}\}$ and learnt jointly by maximizing the log marginal likelihood $\mathcal{L}$ of the targets $\mathbf{y}$, as follows:

$$\log p(\mathbf{y}|\gamma, X) \propto -\left[\mathbf{y}^\top (K_\gamma + \sigma^2 I)^{-1}\mathbf{y} + \log |K_\gamma + \sigma^2 I|\right]. \tag{2}$$

To learn the kernel, the chain rule is applied to compute

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} = \frac{\partial \mathcal{L}}{\partial K_\gamma}\frac{\partial K_\gamma}{\partial \boldsymbol{\theta}} \tag{3}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{\partial \mathcal{L}}{\partial K_\gamma}\frac{\partial K_\gamma}{\partial g(\mathbf{x}, \mathbf{w})}\frac{\partial g(\mathbf{x}, \mathbf{w})}{\partial \mathbf{w}} \tag{4}$$

where the implicit derivative of the log marginal likelihood with respect to the data covariance matrix K_γ is provided by

$$\frac{\partial \mathcal{L}}{\partial K_\gamma} = \frac{1}{2}\left(K_\gamma^{-1}\mathbf{y}\mathbf{y}^\top K_\gamma^{-1} - K_\gamma^{-1}\right).$$

For scalability, a structured kernel interpolation [44] covariance matrix, K_{SKI}, is used instead of K_γ:

$$K_\gamma \approx W K_{U,U} W^\top := K_{\text{SKI}} \tag{5}$$

where U is the set of grid inducing points, $K_{U,U}$ is the kernel matrix between the inducing points, and W is a sparse matrix of the interpolation weights.

2.4. Training Details and Model Hyperparameters

We used the GPyTorch [45] library to implement the model described in the deep kernel learning (DKL) section. To obtain optimal model hyperparameters, we used the Optuna hyperparameter optimization framework [46]. Table 1 contains more information about the model's implementation and its associated hyperparameters.

Table 1. Hyperparameters for the deep kernel model used in our experiments.

Hyperparameter	Options	Optimal Setting
Basis kernel function	RBF	RBF
Grid size	16 to 100	35
Learning rate	1×10^{-5} to 3×10^{-1}	0.0130925
Optimizer	SGD or RMSprop or Adam	RMSprop
Deep neural network (DNN) layers	2 to 7	5
Units in each layer (except the last)	32 to 512	163
Units in the last layer of the DNN	2	2
Activation function	ReLU or LeakyReLU or Tanh	LeakyReLU

With the hyperparameters fixed as shown in the optimal setting column of Table 1, a DKL model was developed using the training set and a five-fold cross-validation scheme, as depicted in Figure 1. A representative model was selected based on the optimal performance during cross-validation. Table S1 in the Supplementary Materials shows the results of cross-validation and which instance of the model was selected. The selected model was then evaluated on the test dataset.

2.5. Performance Evaluation Metrics

To evaluate the performance of the model, we used standard statistical metrics that are commonly used on regression problems. These metrics were the mean squared error (MSE),

root mean squared error (RMSE), coefficient of determination(R^2), and average absolute relative deviation (AARD).

With N samples of data where the measured $\log EC_{50}$ from experiments for sample i is provided by y_i^{exp} and the corresponding prediction from the DKL model by y_i^{pred}, the aforementioned metrics can be obtained as follows:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i^{pred} - y_i^{exp}\right)^2 \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i^{pred} - y_i^{exp}\right)^2} \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i^{pred} - y_i^{exp}\right)^2}{\sum_{i=1}^{N}\left(y_i^{pred} - \bar{y}\right)^2} \tag{8}$$

$$\mathrm{AARD} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i^{pred} - y_i^{exp}}{y_i^{exp}}\right| \tag{9}$$

Note that the term $\bar{y}$ in Equation (8) represents the average measured $\log EC_{50}$ in the dataset.

3. Results and Discussion

In this section, we provide results showing the performance of the DKL model and compare it with GC and ASFC models, two of the state-of-the-art models in this area. These two models, especially ASFC, have been shown to be accurate and reliable in predicting the toxicity of ionic liquids towards the leukemia rat cell line IPC-81. Here, we determine whether DKL can be as accurate and reliable as ASFC.

Table 2 compares the performance of the DKL model with the existing models GC and ASFC. On the 140 ionic liquids used for cross-validation, DKL performs well in all metrics compared to both GC and ASFC. In particular, DKL achieves an RMSE of 0.233, which is about 10% lower than the RMSE achieved by ASFC. The determination coefficient, R^2, achieved by DKL was 0.94, compared to 0.93 for ASFC and 0.924 for GC.

Table 2. Performance comparison of our DKL model with existing models on the training, validation, and full datasets.

Model	Dataset	Samples	AARD%↓	R^2↑	MSE↓	RMSE↓
GC	train + valid	140	11.358	0.924	0.071	0.267
ASFC	train + valid	140	10.898	0.930	0.065	0.256
DKL	train + valid	140	8.756	0.940	0.054	0.233
GC	full	155	-	-	-	-
ASFC	full	155	10.613	0.911	0.086	0.294
DKL	full	155	8.932	0.943	0.052	0.228

Similarly, on the full dataset containing 155 ionic liquids DKL achieved an RMSE of 0.228, compared to 0.294 achieved by ASFC, representing an improvement of about 22%. The coefficient of determination rose from 0.911 for ASFC to 0.943 for DKL.

It is important to note that on both sets of results DKL achieved an RMSE of around 0.23 and an R^2 of about 0.94. This is in contrast to deviations of 0.256–0.294 in RMSE and 0.93–0.911 in R^2 achieved by ASFC, which are slightly larger. The minor deviations in the scores obtained by DKL could mean that the model was not overfitted, and is thus better able to generalize.

The contrast in the performance of ASFC and DKL can further be discerned in Figure 4. The figure shows the sorted absolute errors between experimental and predicted log EC_{50} for ASFC and DKL models on the full dataset of 155 ionic liquids. The area under the absolute error curve associated with DKL is evidently smaller than that of ASFC, revealing a higher predictive accuracy for IL toxicity with the DKL model.

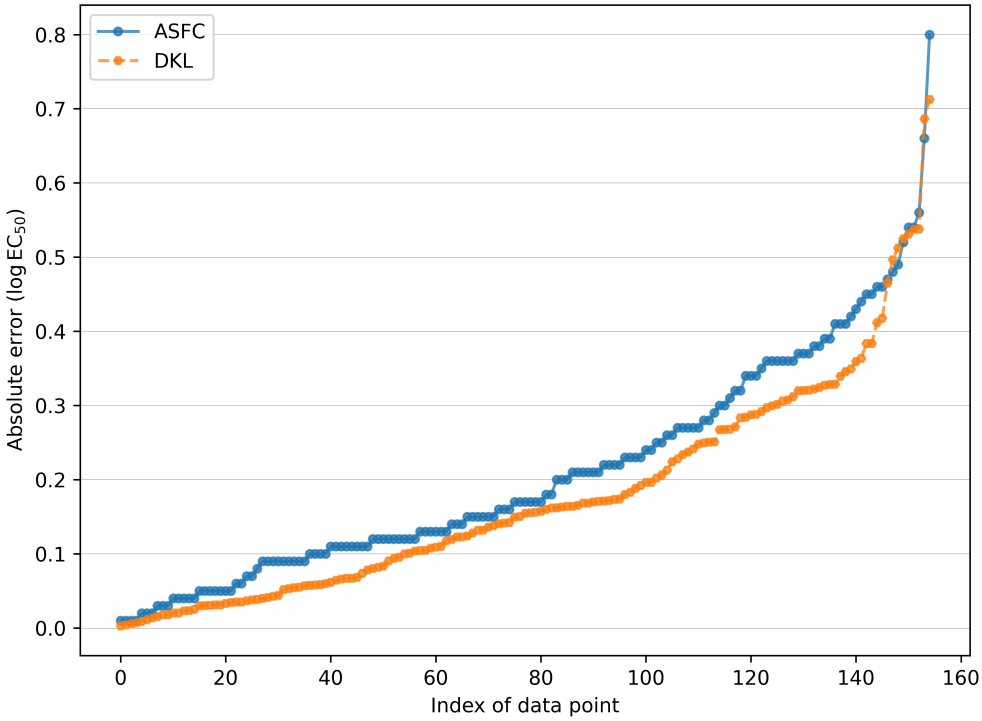

Figure 4. Sorted absolute errors between experimental and predicted log EC_{50} for ASFC and DKL models on the 155 ionic liquids which form the entire dataset.

Because the DKL model is a probabilistic model, it can be used to make predictions for any number of samples while observing the mean predictions and covariances. This information can then be used to determine the uncertainty in the predictions made by the model. Figure 5 shows a comparison between experimental and predicted log EC_{50} for the fifteen ionic liquids that formed the test dataset. It can be observed that the mean predictions made by DKL are close to the experimental log EC_{50} values. This demonstrates that the model learned well and can make authentic predictions. More importantly, we can query the model to show a number of samples that contribute to this prediction, from which we can visually determine the levels of uncertainty in the model. In Figure 5, we show twenty such samples for each of the fifteen predictions.

Figure 6 shows the log EC_{50} values predicted by the DKL model in comparison with the values measured by experiment for the same fifteen ionic liquids used in the test dataset, this time using the indices of the ILs in the dataset as the x-axis variable. From the figure, it can be observed that the model is more uncertain for ILs at index 1, while being more certain about other predictions, such as the prediction at index 0. Such information is important in allowing practitioners or researchers to make decisions about the predictions made by the model. Consider a situation where the chemical structure of the ionic liquid being evaluated is very similar to two other ILs which have very different levels of toxicity, and the latter two were used for modeling. Ideally, the model's uncertainty should be high in order to reflect the varied toxicity levels of the data on which it was modeled. If the uncertainty range enters regions where the toxicity levels are unacceptable, the practitioner may conduct further experiments or gather more information from other sources in order to obtain additional insight about the IL. This extra information would then lead to deciding whether or not to proceed with use of the IL in the intended application.

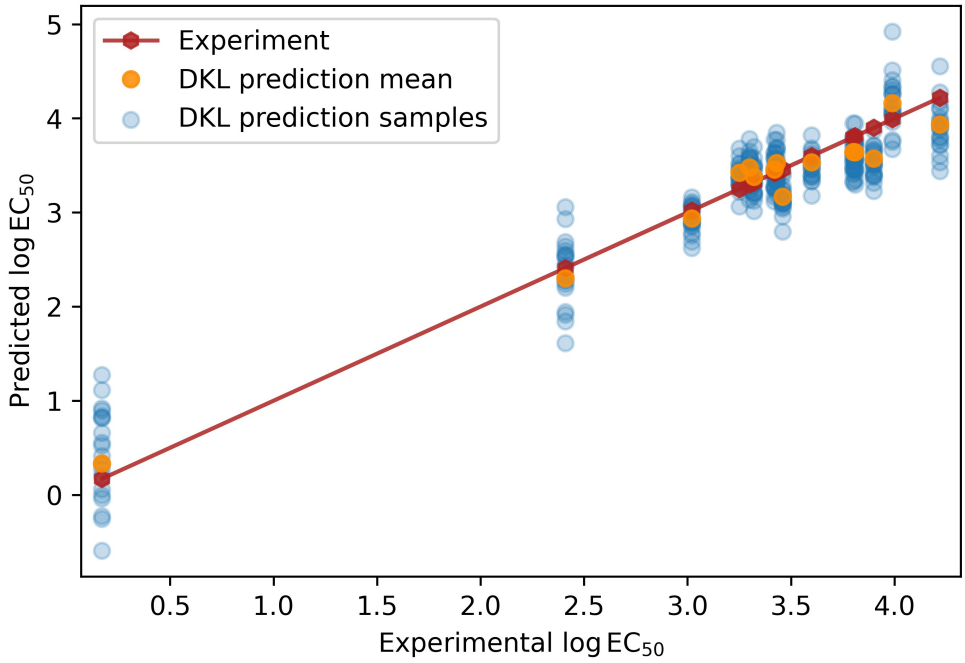

Figure 5. Comparison between experimental and DKL predicted log EC_{50} for the fifteen ionic liquids in the test dataset.

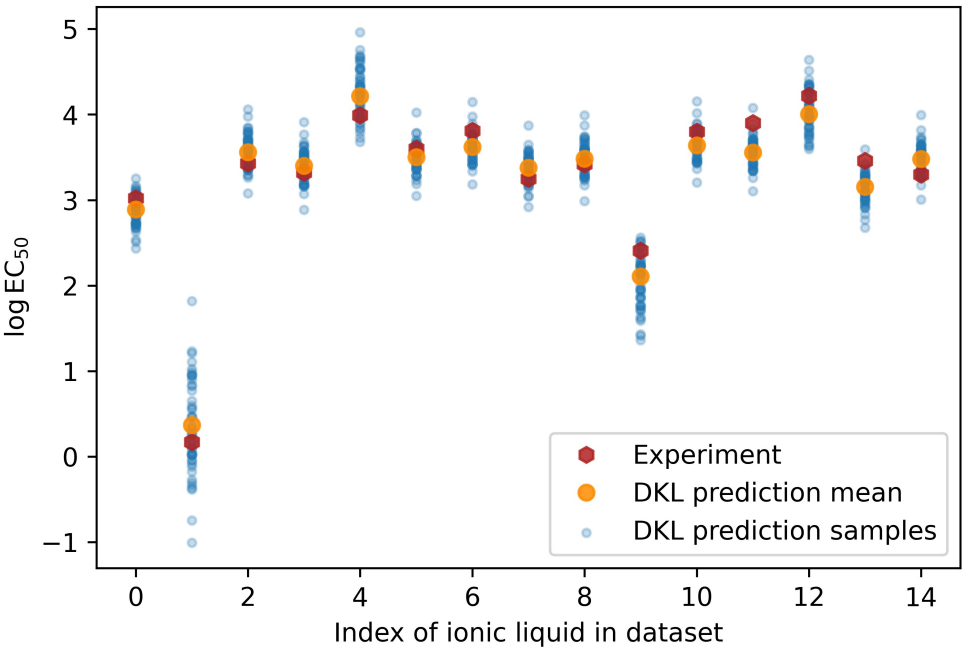

Figure 6. Comparisons between experiment and DKL predicted log EC_{50} for each of the fifteen ionic liquids forming the test dataset. For each DKL prediction, we drew samples that contribute to the mean prediction.

3.1. Applicability Domain

As per Organisation for Economic Co-operation and Development (OECD) principles which stipulate that Quantitative Structure–Activity Relationship (QSAR) prediction models should have well-defined applicability domains (AD), we performed an AD analysis for this study. We used the standardization technique (ST) proposed by Roy et al. [47]. In an ideal situation, data are distributed such that 99.7% of the population falls within the range mean ± 3 standard deviations (SD). In this context, this implies that mean ± 3SD

represents the zone to which the majority of the ionic liquids in the training set belong. Any ionic liquids appearing outside this region are considered to be different from the rest of the ionic liquids.

In ST, a descriptor column is standardized based on the corresponding mean and standard deviation for the training set only. If the ensuing standardized value for a descriptor of a particular ionic liquid is more than 3.0, then the ionic liquid is considered an outlier if it is in the training set, and is considered outside the AD if it is part of the test set [48]. The applicability domain section in the Supplementary Material provides a full description of the ST algorithm.

The distribution map of the applicability domain is shown in Figure 7. The coverage of the test set in the applicability domain using the ST shows that all but one ionic liquid fell outside the AD. Similarly, in the training set, three of the 140 ionic liquids were considered outliers. This means that 93% and 98% of the ionic liquids in the test and training sets, respectively, fall within the AD.

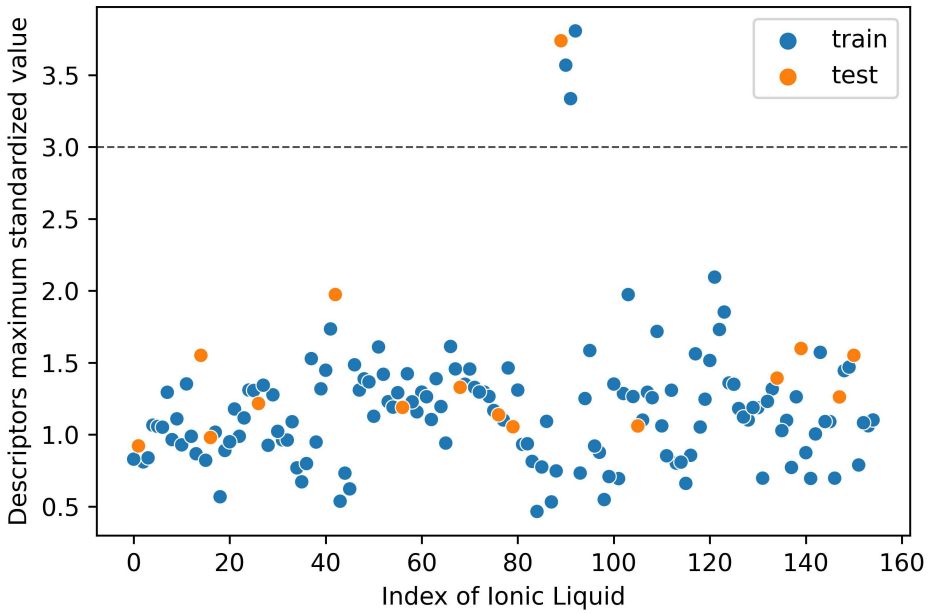

Figure 7. Applicability domain defined in this study. One ionic liquid in the test set is outside the AD, and three ILs in the training set are outliers as their corresponding descriptors' maximum standardized values are greater than the 3.0 threshold.

Our DKL model uses the "mixtures out" validation protocol. To a large extent, this protocol estimates the ability of models to predict new combinations of anions and cations. This may provide overoptimistic results, as described elsewhere [49,50]. This applies to the ASFC model with which we are comparing DKL in this study as well. There exist more rigorous validation protocols, such as "components validation", which can test the model's prediction of new types of ions. By design, components validation is more similar to real-life situations [49,51]. Thus, replacing mixtures validation with components validation in our workflow may minimize the level of optimistic results, if any.

3.2. Prediction Web Server

A web server that encapsulates the DKL model was built. The tool accepts SMILES strings as input for the ionic liquids and provides results in both tabular and interactive visualization formats. The server is publicly available at http://nsclbio.jbnu.ac.kr/tools/iltox/, accessed on 8 April 2022.

4. Conclusions

Currently available data do not show that ionic liquids are environmentally safe chemicals; as such, their toxicity risk has to be evaluated in order to ensure their safe use in a wide range of applications. In this work, we have presented a probabilistic deep learning model that can be used to predict the toxicity of ionic liquids towards the leukemia rat cell-line (IPC-81) reliably and accurately. The model pipeline requires little or no expert domain knowledge in the generation of features to be used for subsequent predictions. In addition, all predictors are generated using open source cheminformatics tools. In addition, because the model is embedded with a Gaussian process it has the inherent capability to attach a level of uncertainty to each prediction it makes. As the dataset used in this work was generated from experimental measurements in which inconsistencies are, at the very least, unavoidable, the uncertainty associated with these data had to be addressed. In that respect, the results obtained here indicate that the presented probabilistic deep learning model represents a good choice. Furthermore, the probabilistic nature of the model means that it provides vital information with which users can interpret prediction results and gain insight about both the data and the model. Finally, based on this model we developed a web-based tool which can be used to make predictions. This tool is freely available on our project website.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/ijms23095258/s1.

Author Contributions: Conceptualization, M.C.; Methodology, M.C.; Software, M.C. and H.T.; Validation, M.C., H.T. and K.T.C.; Resources, K.T.C.; Writing—Original Draft, M.C.; Writing—Review & Editing, H.T. and K.T.C.; Visualization, M.C. and H.T.; Supervision, K.T.C.; Project Administration, H.T. and K.T.C.; Funding Acquisition, K.T.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2020R1A2C2005612).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this work can be found in the attached Supplementary Material and can be downloaded from the project web page at http://nsclbio.jbnu.ac.kr/tools/iltox/, accessed on 8 April 2022.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

EC	Effective Concentration
GC	Group Contribution
DKL	Deep Kernel Learning
MSE	Mean Squared Error
SVM	Support Vector Machine
AARD	Average Absolute Relative Deviation
ASFC	Atom Surface Fragment Contribution
OECD	Organisation for Economic Co-operation and Development
RMSE	Root Mean Squared Error
QSAR	Quantitative Structure–Activity Relationship
SMILES	Simplified Molecular-Input Line Entry System

References

1. Armand, M.; Tarascon, J.M. Building better batteries. *Nature* **2008**, *451*, 652–657. [CrossRef] [PubMed]
2. Armand, M.; Endres, F.; MacFarlane, D.R.; Ohno, H.; Scrosati, B. Ionic-liquid materials for the electrochemical challenges of the future. In *Materials for Sustainable Energy*; Nature Publishing Group: London, UK, 2011; pp. 129–137.
3. Magina, S.; Barros-Timmons, A.; Ventura, S.P.; Evtuguin, D.V. Evaluating the hazardous impact of ionic liquids–challenges and opportunities. *J. Hazard. Mater.* **2021**, *412*, 125215. [CrossRef] [PubMed]
4. Gonçalves, A.R.; Paredes, X.; Cristino, A.; Santos, F.; Queirós, C.S. Ionic liquids—A review of their toxicity to living organisms. *Int. J. Mol. Sci.* **2021**, *22*, 5612. [CrossRef] [PubMed]
5. Zhao, D.; Fei, Z.; Geldbach, T.J.; Scopelliti, R.; Dyson, P.J. Nitrile-functionalized pyridinium ionic liquids: Synthesis, characterization, and their application in carbon- carbon coupling reactions. *J. Am. Chem. Soc.* **2004**, *126*, 15876–15882. [CrossRef]
6. Ta, L.; Axelsson, A.; Bijl, J.; Haukka, M.; Sundén, H. Ionic Liquids as Precatalysts in the Highly Stereoselective Conjugate Addition of α, β-Unsaturated Aldehydes to Chalcones. *Chem. Eur. J.* **2014**, *20*, 13889–13893. [CrossRef]
7. Stoimenovski, J.; MacFarlane, D.R.; Bica, K.; Rogers, R.D. Crystalline vs. ionic liquid salt forms of active pharmaceutical ingredients: A position paper. *Pharm. Res.* **2010**, *27*, 521–526. [CrossRef]
8. Postleb, F.; Stefanik, D.; Seifert, H.; Giernoth, R. Bionic liquids: Imidazolium-based ionic liquids with antimicrobial activity. *Z. Naturforschung B* **2013**, *68*, 1123–1128. [CrossRef]
9. Swatloski, R.P.; Spear, S.K.; Holbrey, J.D.; Rogers, R.D. Dissolution of cellose with ionic liquids. *J. Am. Chem. Soc.* **2002**, *124*, 4974–4975. [CrossRef]
10. Rao, C.J.; Venkatesan, K.; Nagarajan, K.; Srinivasan, T.; Rao, P.V. Electrochemical behavior of europium (III) in N-butyl-N-methylpyrrolidinium bis (trifluoromethylsulfonyl) imide. *Electrochim. Acta* **2009**, *54*, 4718–4725.
11. Rao, C.J.; Venkatesan, K.; Nagarajan, K.; Srinivasan, T.; Rao, P.V. Electrodeposition of metallic uranium at near ambient conditions from room temperature ionic liquid. *J. Nucl. Mater.* **2011**, *408*, 25–29.
12. Wu, B.; Reddy, R.G.; Rogers, R.D. Novel ionic liquid thermal storage for solar thermal electric power systems. In *International Solar Energy Conference*; American Society of Mechanical Engineers: Washington, DC, USA, 2001; Volume 16702, pp. 445–451.
13. Cho, C.W.; Pham, T.P.T.; Zhao, Y.; Stolte, S.; Yun, Y.S. Review of the toxic effects of ionic liquids. *Sci. Total Environ.* **2021**, *786*, 147309. [CrossRef] [PubMed]
14. Torrecilla, J.S.; García, J.; Rojo, E.; Rodríguez, F. Estimation of toxicity of ionic liquids in Leukemia Rat Cell Line and Acetylcholinesterase enzyme by principal component analysis, neural networks and multiple lineal regressions. *J. Hazard. Mater.* **2009**, *164*, 182–194. [CrossRef] [PubMed]
15. Samorì, C.; Pasteris, A.; Galletti, P.; Tagliavini, E. Acute toxicity of oxygenated and nonoxygenated imidazolium-based ionic liquids to Daphnia magna and Vibrio fischeri. *Environ. Toxicol. Chem. Int. J.* **2007**, *26*, 2379–2382. [CrossRef] [PubMed]
16. Latała, A.; Nędzi, M.; Stepnowski, P. Toxicity of imidazolium and pyridinium based ionic liquids towards algae. Bacillaria paxillifer (a microphytobenthic diatom) and Geitlerinema amphibium (a microphytobenthic blue green alga). *Green Chem.* **2009**, *11*, 1371–1376. [CrossRef]
17. Wang, Z.; Song, Z.; Zhou, T. Machine learning for ionic liquid toxicity prediction. *Processes* **2021**, *9*, 65. [CrossRef]
18. Stolte, S.; Matzke, M.; Arning, J.; Böschen, A.; Pitner, W.R.; Welz-Biermann, U.; Jastorff, B.; Ranke, J. Effects of different head groups and functionalised side chains on the aquatic toxicity of ionic liquids. *Green Chem.* **2007**, *9*, 1170–1179. [CrossRef]
19. Ranke, J.; Stolte, S.; Störmann, R.; Arning, J.; Jastorff, B. Design of sustainable chemical products the example of ionic liquids. *Chem. Rev.* **2007**, *107*, 2183–2206. [CrossRef]
20. Zhao, Y.; Zhao, J.; Huang, Y.; Zhou, Q.; Zhang, X.; Zhang, S. Toxicity of ionic liquids: Database and prediction via quantitative structure—Activity relationship method. *J. Hazard. Mater.* **2014**, *278*, 320–329. [CrossRef]
21. Cho, C.W.; Stolte, S.; Yun, Y.S. Comprehensive approach for predicting toxicological effects of ionic liquids on several biological systems using unified descriptors. *Sci. Rep.* **2016**, *6*, 33403. [CrossRef]
22. Sosnowska, A.; Grzonkowska, M.; Puzyn, T. Global versus local QSAR models for predicting ionic liquids toxicity against IPC-81 leukemia rat cell line: The predictive ability. *J. Mol. Liq.* **2017**, *231*, 333–340. [CrossRef]
23. Cao, L.; Zhu, P.; Zhao, Y.; Zhao, J. Using machine learning and quantum chemistry descriptors to predict the toxicity of ionic liquids. *J. Hazard. Mater.* **2018**, *352*, 17–26. [CrossRef] [PubMed]
24. Kang, X.; Chen, Z.; Zhao, Y. Assessing the ecotoxicity of ionic liquids on Vibrio fischeri using electrostatic potential descriptors. *J. Hazard. Mater.* **2020**, *397*, 122761. [CrossRef] [PubMed]
25. Wu, T.; Li, W.; Chen, M.; Zhou, Y.; Zhang, Q. Estimation of Ionic Liquids Toxicity against Leukemia Rat Cell Line IPC-81 based on the Empirical-like Models using Intuitive and Explainable Fingerprint Descriptors. *Mol. Inform.* **2020**, *39*, 2000102. [CrossRef] [PubMed]
26. Wang, Z.; Su, Y.; Jin, S.; Shen, W.; Ren, J.; Zhang, X.; Clark, J.H. A novel unambiguous strategy of molecular feature extraction in machine learning assisted predictive models for environmental properties. *Green Chem.* **2020**, *22*, 3867–3876. [CrossRef]
27. Landrum, G. RDKit: Open-Source Cheminformatics Software. 2016. Available online: https://www.rdkit.org/ (accessed on 8 April 2022).
28. Lin, S.T.; Sandler, S.I. Henry's law constant of organic compounds in water from a group contribution model with multipole corrections. *Chem. Eng. Sci.* **2002**, *57*, 2727–2733. [CrossRef]

29. Sedlbauer, J.; Bergin, G.; Majer, V. Group contribution method for Henry's Law constant of aqueous hydrocarbons. *AIChE J.* **2002**, *48*, 2936–2959. [CrossRef]
30. Huang, Y.; Dong, H.; Zhang, X.; Li, C.; Zhang, S. A new fragment contribution-corresponding states method for physicochemical properties prediction of ionic liquids. *AIChE J.* **2013**, *59*, 1348–1359. [CrossRef]
31. Razdan, N.K.; Koshy, D.M.; Prausnitz, J.M. Henry's constants of persistent organic pollutants by a group-contribution method based on scaled-particle theory. *Environ. Sci. Technol.* **2017**, *51*, 12466–12472. [CrossRef]
32. Peng, D.; Picchioni, F. Prediction of toxicity of Ionic Liquids based on GC-COSMO method. *J. Hazard. Mater.* **2020**, *398*, 122964. [CrossRef]
33. Kang, X.; Zhao, Y.; Chen, Z. Atom surface fragment contribution method for predicting the toxicity of ionic liquids. *J. Hazard. Mater.* **2022**, *421*, 126705. [CrossRef]
34. Mu, T.; Rarey, J.; Gmehling, J. Group contribution prediction of surface charge density profiles for COSMO-RS (Ol). *AIChE J.* **2007**, *53*, 3231–3240. [CrossRef]
35. Abramenko, N.; Kustov, L.; Metelytsia, L.; Kovalishyn, V.; Tetko, I.; Peijnenburg, W. A review of recent advances towards the development of QSAR models for toxicity assessment of ionic liquids. *J. Hazard. Mater.* **2020**, *384*, 121429. [CrossRef] [PubMed]
36. Hossain, M.I.; Samir, B.B.; El-Harbawi, M.; Masri, A.N.; Mutalib, M.A.; Hefter, G.; Yin, C.Y. Development of a novel mathematical model using a group contribution method for prediction of ionic liquid toxicities. *Chemosphere* **2011**, *85*, 990–994. [CrossRef] [PubMed]
37. Hüllermeier, E.; Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Mach. Learn.* **2021**, *110*, 457–506. [CrossRef]
38. Zhang, S.; Sun, N.; He, X.; Lu, X.; Zhang, X. Physical properties of ionic liquids: Database and evaluation. *J. Phys. Chem. Ref. Data* **2006**, *35*, 1475–1517. [CrossRef]
39. Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27–35. [CrossRef]
40. Irwin, J.J.; Shoichet, B.K. ZINC- a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182. [CrossRef]
41. Irwin, J.J.; Sterling, T.; Mysinger, M.M.; Bolstad, E.S.; Coleman, R.G. ZINC: A free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757–1768. [CrossRef]
42. Gaulton, A.; Bellis, L.J.; Bento, A.P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107. [CrossRef]
43. Wilson, A.G.; Hu, Z.; Salakhutdinov, R.; Xing, E.P. Deep kernel learning. In Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, Cadiz, Spain, 9–11 May 2016; pp. 370–378.
44. Wilson, A.; Nickisch, H. Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 7–9 July 2015; pp. 1775–1784.
45. Gardner, J.; Pleiss, G.; Weinberger, K.Q.; Bindel, D.; Wilson, A.G. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.
46. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631.
47. Roy, K.; Kar, S.; Ambure, P. On a simple approach for determining applicability domain of QSAR models. *Chemom. Intell. Lab. Syst.* **2015**, *145*, 22–29. [CrossRef]
48. Kar, S.; Roy, K.; Leszczynski, J. Applicability domain: A step toward confident predictions and decidability for QSAR modeling. In *Computational Toxicology*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 141–169.
49. Makarov, D.; Fadeeva, Y.A.; Shmukler, L.; Tetko, I. Beware of proper validation of models for ionic Liquids! *J. Mol. Liq.* **2021**, *344*, 117722. [CrossRef]
50. Muratov, E.N.; Varlamova, E.V.; Artemenko, A.G.; Polishchuk, P.G.; Kuz'min, V.E. Existing and developing approaches for QSAR analysis of mixtures. *Mol. Informat.* **2012**, *31*, 202–221. [CrossRef] [PubMed]
51. Oprisiu, I.; Novotarskyi, S.; Tetko, I.V. Modeling of non-additive mixture properties using the Online CHEmical database and Modeling environment (OCHEM). *J. Cheminform.* **2013**, *5*, 4. [CrossRef] [PubMed]