# Transfer Learning Graph Representations of Molecules for pKa, $^{13}$C-NMR, and Solubility

A. M. El-Samman*[a], S. De Castro[a], B. Morton[a], and S. De Baerdemacker[a,b]

[a] University of New Brunswick, Department of Chemistry. 30 Dineen Dr, Fredericton, Canada. E-mail: aelsamma@unb.ca

[b] University of New Brunswick, Department of Mathematics and Statistics. 30 Dineen Dr, Fredericton, Canada.

**Abstract**

We explore transfer learning models from a pre-trained graph convoluntional neural network representation of molecules, obtained from SchNet,[1] to predict $^{13}$C-NMR, pKa, and log$S$ solubility. SchNet learns a graph representation of a molecule by associating each atom with an "embedding vector" and interacts the atom-embeddings with each other by leveraging graph-convolutional filters on their interatomic distances. We pre-trained SchNet on molecular energy and demonstrate that the pre-trained atomistic embeddings can then be used as a transferable representation for a wide array of properties. On the one hand, for atomic properties such as micro-pK1 and $^{13}$C-NMR, we investigate two models, one linear and one neural net, that inputs pre-trained atom-embeddings of a particular atom (e.g. carbon) and predicts a local property (e.g. $^{13}$C-NMR). On the other hand, for molecular properties such as solubility, a size-extensive graph model is built using the embeddings of all atoms in the molecule as input. For all cases, qualitatively correct predictions are made with relatively little training data ($< 1000$ training points), showcasing the ease with which pre-trained embeddings pick up on important chemical patterns. The proposed models successfully capture well-understood trends of pK1 and solubility. This study advances our understanding of current neural net graph representations and their capacity for transfer learning applications in chemistry.

***Keywords***— Machine Learning, Molecular Representations, Transferable Representations, Transfer Learning, Molecular Descriptors, Graph Neural Networks, Graph Descriptors, Embeddings, Electronic Structure, Chemical Properties, pKa, NMR, log$S$, solubility

# Introduction

Prediction of protonation constants (pKa), nuclear magnetic resonance ($^{13}$C-NMR), and solubility (log$S$) has traditionally relied on either physics-based theoretical methods[2–14] or empirical approaches.[15–22] While these methods have provided valuable insights into molecular behavior, they often involve complex calculations and may struggle to handle larger molecules accurately. In recent years, machine learning approaches[23–33] have emerged as promising alternatives for predicting molecular properties. These approaches can generally be divided into two categories: descriptor-based and end-to-end based.

Descriptor-based methods[32–38] involve deriving numerical descriptors or features from the molecular structure that encapsulate important characteristics about the molecule. Common descriptors include molecular fingerprints,[39;40] which encode information about molecular substructures, and physiochemical properties[37] such as molecular weight, polarizability,[41] and hydrogen bonding potential.[42] Descriptor-based models can be trained using a diverse set of molecular properties to learn the relationships between these descriptors and the target properties, such as pKa and $^{13}$C-NMR chemical shifts. While descriptor-based approaches

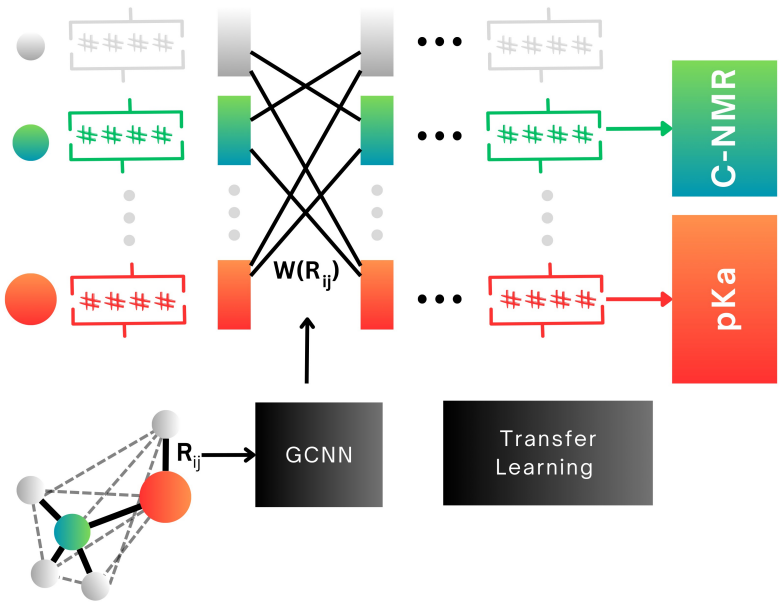



Figure 1: Schematic diagram of the GCNN SchNet [1;66] and the subsequent transfer learning model. Embedding vectors are trained on total energies of the QM9 [84] dataset using SchNet, extracted and used as input features for transfer learning on pKa, $^{13}$C-NMR, etc.

have shown success in certain applications, they may face challenges in capturing more intricate molecular interactions and non-linear relationships limiting their performance for complex molecular systems. While results are controversial, [43;44] there is plenty of support that shows that end-to-end graph-based predictions can outperform descriptor-based ones. [45–50] Wu et. al. reported on MoleculeNet, a large benchmark for molecular machine learning tasks, and the evaluation results illustrated that graph-based methods outperformed descriptor-based methods on most datasets. [50] Graph-based methods are also exempt from rigorous feature engineering, and do not need to be fine-tuned to suit a specific purpose. [46;47]

In end-to-end graph-based approaches, [51–60] molecules are represented as graphs, with atoms as nodes and connections as edges, adhering more closely to the chemically intuitive representation of a molecule. Approaches such as graph convolutional neural networks (GCNNs) [1;59–65] have been applied successfully to predict molecular properties. A schematic representation of SchNet [1;66], a typical graph convolutional neural network architecture is shown on the left part of Figure 1. The central objects of molecular GCNNs, particularly in SchNet, are the atomwise embedding vectors, high-dimensional vectors, that are associated to the nodes in the graph. In the end-to-end approach, no prior feature engineering is used to pre-design these atomwise embedding vectors. They dynamically learn the relevant features of the molecule during the training process via a concatenation of interactions with the other atoms' embedding vectors.

Although GCNNs have been originally designed to predict extensive molecular properties, such as potential energy [1;62;66–68], their atomwise architecture makes them tailor made to predict atom-based properties

of molecules, such as pKa and NMR shifts.[51–54;60;69] For instance, Xiong et al.[54] compiled a large-scale pKa dataset containing 16,595 compounds with 17,489 pKa values for use in their Graph-pKa model, giving a mean absolute error of around 0.55 on their macro-pKa test set. In another work, Guan et al. designed a graph-based neural net to predict $^{1}$H- and $^{13}$C-NMR chemical shits then tested it on the CHESHIRE dataset to give an error of 1.23 ppm on chemical shifts, which is comparable to DFT functional accuracy. Their algorithm was trained on 100,000 $^{1}$H- and $^{13}$C-NMR chemical shifts computed from DFT-optimized structures, and then retrained on a smaller set of experimental NMR shifts to improve accuracy.

A common theme of graph-based neural nets is the requirement for large databases, this was a major caveat found in the results of the MoleculeNet benchmark database which was used to compare performance between various graph-based and descriptor-based algorithms.[50] It was found that graph-based neural nets struggle to handle smaller datasets and are outperformed by traditional descriptor-based methods in these situtations. This is indeed a problem, as experimental determination of many properties in chemistry is often a very time-consuming task, requiring expert knowledge and specialized instruments, making molecular datasets often too small for graph-based algorithms.

Transfer learning is known to be a great remedy for this problem.[70–78] Transfer learning leverages generalizable knowledge already contained within pre-trained graph-based molecular representations to retrain (and accurately perform) on smaller datasets. This can be a great benefit when data is scarce as the ipnut representation will come equipped with important features gained from the pre-training. Transfer learning has been done for graph-based representations, even in the context of pKa, NMR and solubility predictions.[53;55;69;79;80] However, for each of these cases, it has been performed in a quasi-transfer way, whereby an algorithm already trained on one of the properties ($\log S$, NMR, or pKa) is retrained on a smaller and more accurate dataset of that same property, usually an experimental database, to improve accuracy. For instance, Vermeire and Green,[80] train their graph-based neural network on computed solvation energies and then retrain their network on a smaller dataset of experimental solvation energies to obtain better accuracy.

We test the limits of the transfer learning hypothesis in a more drastic way. We showed in a previous work,[81] that the atomwise embedding vectors of the GCNN SchNet trained on molecular energies learns crucial atom-based information about the chemical neighborhood that the atom resides in. This remarkable feature of the trained model gives confidence that it can be used for transfer learning purposes towards a much wider range of chemical properties. The ultimate goal is to reach for a general representation that can handle just about any property, akin to how all molecular properties can be derived from the wave function of a molecule. In machine learning statistical representations, this is referred to as "task diversity" and it is recently being recognized as a general property of neural network models.[70;82;83] Roughly put, it is a statistical measure of how many samples are required to learn a representation (shared across tasks) and use it to improve prediction on a new task. It has been used to derive statistical guarantees about the size of samples required to transfer train on a new task.[70]

Fortunately, graph-based representations are often parsed in an atomwise manner making them directly suitable for transfer learning on atomic/local properties of chemistry. This means that we do not necessarily need to retrain an entire GCNN on a new molecular property but can transfer learn using simpler procedures (linear regression, dense neural nets) directly on the atomistic embeddings. This opens up the opportunity to use more interpretable and informative procedures. pKa and NMR are local properties and are therefore well-suited to interface with the atom-based embedding vectors from SchNet.

However, it is also interesting to investigate transfer learning towards other extensive molecular properties, such as solubility. As the internal model of the GCNN SchNet is atom-based, it builds a model of the energy of the molecule as an extensive sum of atomistic energy contributions. In the same vein for solubility, there is the potential to obtain a learned atomistic interpretation predicting each atom's contributions to solubility using the log$S$ parameter.

# Methodology

## Data

To obtain the embedding vectors, our input representation for transfer learning, we pre-trained a SchNet neural network with six interaction layers, each with 128 atom basis and 128 convolutional filters. The network employed 50 Gaussians with an interaction cutoff of 50 Å to model the interatomic interactions accurately. The QM9 dataset [84] was used to pre-train SchNet. This is a set of 134K small-sized organic molecules ($\sim$ 5-10 Å in size) with optimized conformations all computed using the B3LYP/6-31G(d,p) level of density-functional theory. We trained on 100k molecules with total electronic energy at 0K as the target property. And additional 10,000 data points were used for validation during the training process. The rest of the set (20,000) was leftover for testing. The starting representation for transfer learning achieves a MAE of 0.2 meV on the training set's molecular energy, and 1 meV on the testing set's molecular energy. Note that more efficient GCNN training algorithms employ cutoff distances that are significantly shorter which allowed to efficiently form neighbor interactions. This avenue was not chosen in our dataset, having such a large cutoff distance was purposeful to maintain a global representation of molecules in the embedding vectors. The extracted embeddings for QM9 molecules (and trained model) can be found at [85].

For pKa transfer models, we curated a dataset of 601 clean data points from the high-confidence IUPAC pK1 values. [86] The selected molecules had high-confidence experimental pK1 values digitized from well-established experimental datasets of pKa. [87–89] Our dataset includes only those that are labeled "Reliable," only first dissociation constants pK1s, and only those experimentally determined in the range of 20 °C to 30 °C. This small but high-quality data set allowed us to transfer train our model effectively from the embedding representation, and accurately predict pKa values.

In the case of $^{13}$C-NMR transfer models, there is a lack of proprietary-free $^{13}$C-NMR databases. To

the best of our knowledge, NMRShiftDB2[90–92], and recently the NP-MRD database[93], are the only well-established large open-access $^{13}$C-NMR databases. NMRShiftDB2 also allows users to input spectra, is peer-reviewed by a board of reviewers, and also allows for $^{13}$C-NMR prediction. We used the NMRShiftDB2 model on a subset of QM9 molecules (200 molecules). The model uses Hierarchically Ordered Spherical Environment (HOSE) codes to describe atomic neighborhoods,[94] a molecular descriptor that uses concentric spheres to describe neighborhoods around atoms. Two atoms having the same neighborhoods will have the same HOSE code. HOSE codes are trained on experimental $^{13}$C-NMR data. We could not gain access to the experimental NMR data from NMRShiftDB2 as the molecular files are not ordered according to $^{13}$C-NMR tabular values, but rather according to a labeled 2D sketch of the molecule, indexed differently from the xyz file. This makes the available experimental NMR data relatively inaccessible to xyz-based ML algorithms. While our approach provides an indirect predictor of experimental $^{13}$C-NMR, it still provides a proof-of-concept of task diversity, in other words, transfer learning from GCNNs that pre-trained on energy. We employed the first 200 molecules of the QM9 dataset and the NMRShiftDB2 model to obtain our $^{13}$C-NMR targets.

For the log$S$ training data, we accessed the Natural Products Magnetic Resonance Database (NP-MRD).[93] This giant database contains ∼100,000 natural products (of 20-40 atoms in size, notably much larger than QM9 molecules) with many properties including log$S$ solubility measurements. The database also includes NMR chemical shifts, however, we found that it exhibits the same data inaccessibility issues as the NMRShiftDB2 database. From NP-MRD, we curated our own small dataset of 800 molecules that contain only elements found in the QM9 database (H,C,N,O,F). These log$S$ values are to be targets of a size-extensive machine learning algorithm that learns them from the whole molecule's atomistic embeddings.

## Statistical Learning Methods

First, transfer learning was tested for atomistic properties such as pKa and NMR shifts. To maximize interpretability of our models, We used a "bottom-up" approach to transfer learning, where we started with simpler models and moved on to more complex. For the sake of self-containdedness, we begin along the lines of previous work,[81] where we used a simple linear regression model to map atom-embeddings to the target of interest. The selected atom type depended on the target property, for $^{13}$C-NMR a carbon was the obvious choice. For pKa, we could have chosen any atom near the deprotonation site such as the hydrogen itself or an oxygen on the adjacent site. Since, our dataset contained mainly oxygen-type acids (such as carboxylic acids and alcohols) it was appropriate to choose the oxygen atom embedding for the study, though we expect the same results from hydrogen embeddings.

A second transfer learning model used a feedforward neural net to map atomistic embeddings to atomistic properties. This neural net was made up of two layers (with "ReLu" activation in between), and a final third linear layer to predict the target property. 200 nodes were used per layer. For each method the data

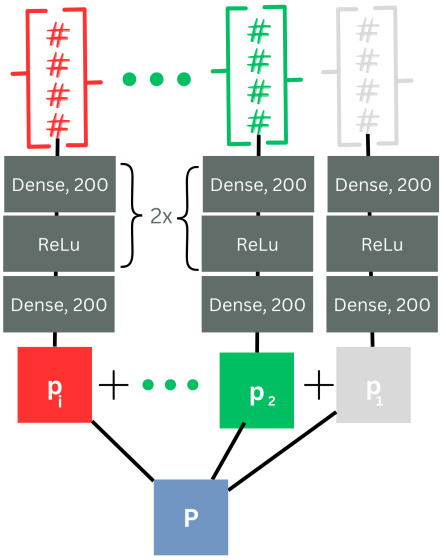



Figure 2: Architecture of the atomwise neural network used to predict atomwise contributions from embedding vectors and pooled to total molecular property.

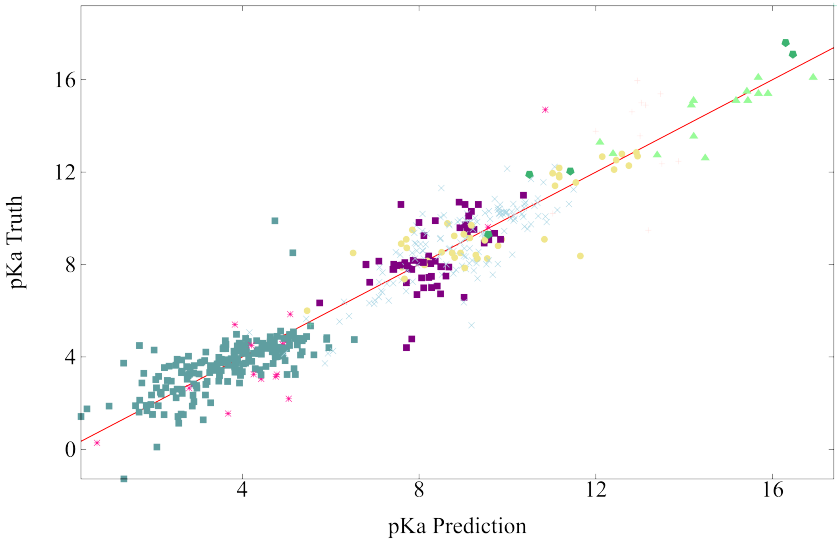
was randomly split into training, validation, and testing data sets using a 6:2:2 ratio on the total data set. The neural net was optimized using the Adam optimizer at a learning rate of $1 \times 10^{-2}$.

Finally for molecular properties, such as log$S$, we designed a third size-extensive transfer neural net. This neural net takes in atomistic embeddings and predicts atomwise contributions to the log$S$ while training on the total log$S$ property. Each atomwise neural net has two layers (with "ReLu" activation in between), and a final third linear layer to predict the atomwise contribution. 200 nodes were used per layer for each atomistic neural net. A diagram of the neural net is shown in Figure 2. Again, the data was randomly split into training, validation, and testing datasets using a 6:2:2 ratio on the total data set. The neural net was optimized using the Adam optimizer at a learning rate of $1 \times 10^{-2}$.

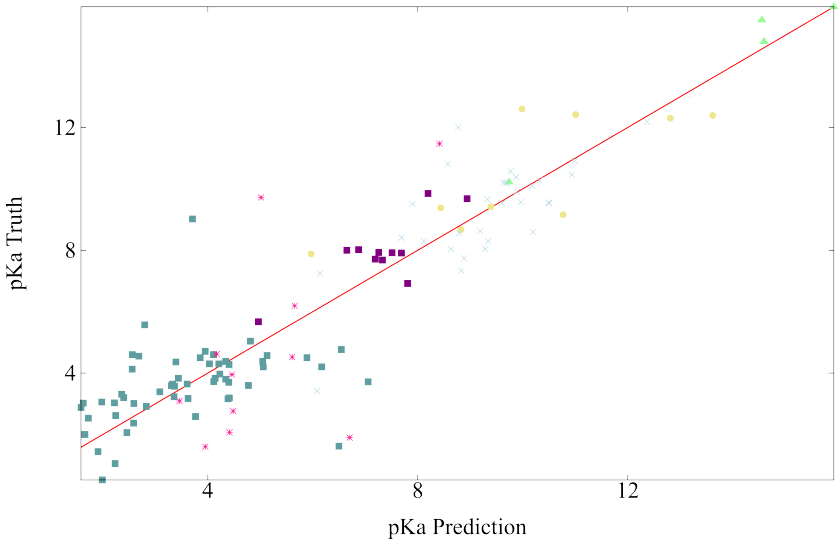
# Results & Discussion

## Linear Transfer Learning – Embeddings to Atomistic pKa/NMR

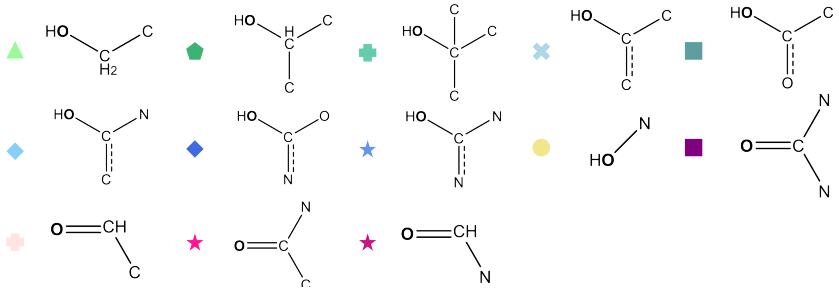
First, we show the results of the linear model between embeddings and pKa/NMR. Figure 3a and 3b show the fit and test results of the pKa model. The predictions of the models are labeled according to oxygen-centric moieties, label key found in Figure 3c. 481 data points were used for training/validation and 120 data points for testing. The linear model gave a RMSE of 1.02 pKa units and 1.44 pKa units for the fit and testing datasets, respectively. For $^{13}$C-NMR, Figure 4a and 4b show the linear fit and test results of the model which are labeled according to carbon-centric moieties, label key found in Figures 4c. The linear model gave a RMSE of 11.71 ppm and 15.09 ppm on training and testing datasets, respectively.




(a)

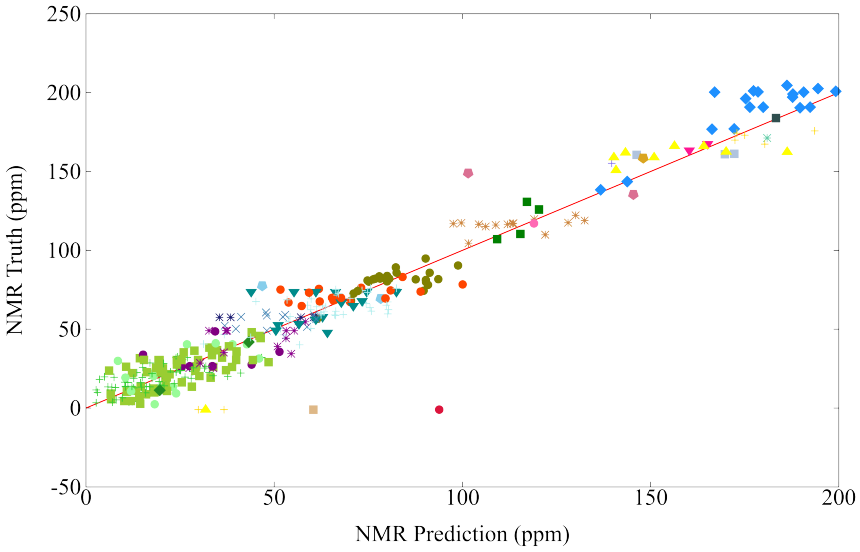



(b)

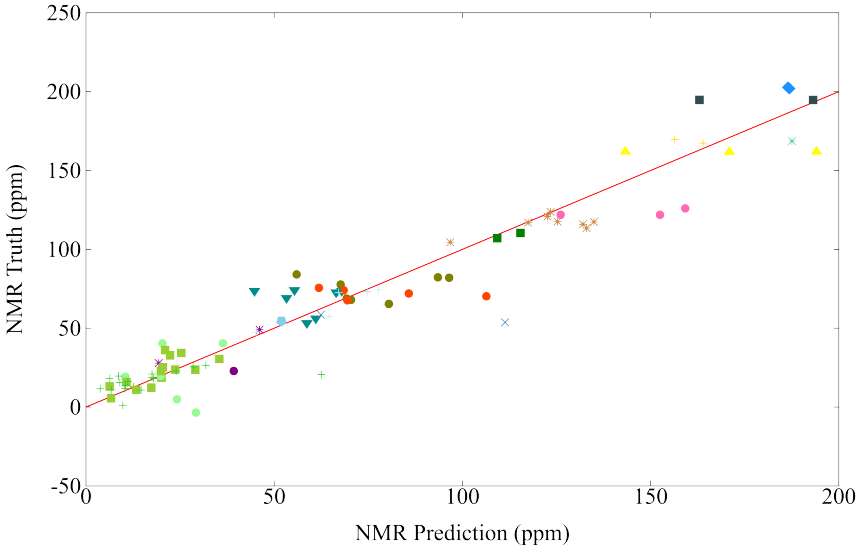


(c)

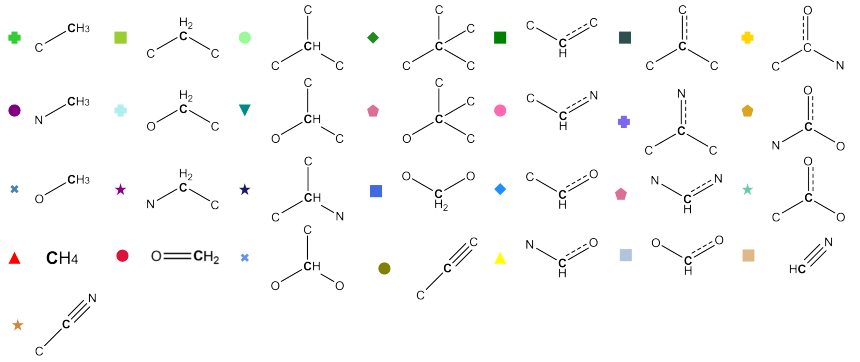
Figure 3: Predictions vs ground truth values for (a) training data (481 datapoints) and (b) test data (121 datapoints) from linear regression for pK1 from oxygen embeddings. Oxygen centered chemical environment labels are given in (c).




(a)




(b)




(c)

Figure 4: Predictions vs ground truth values for (a) training data (400 datapoints) and (b) test data (100 datapoints) from linear regression for $^{13}$C-NMR from carbon embeddings. Carbon centered chemical environment labels are given in (c).

The linear regression results for pK1/NMR are considerably accurate given the small size of the fitting data used. In comparison to other studies, e.g. Mayer et al.[53] who achieved a RMSE of 0.82 and 0.93 on their pKa test sets after training their GCNN on 714,906 microstate pKa values from the CheMBL database,[95] and further supplementing the training with transfer learning on 5,994 experimental pKa values. The pre-transfer learning model of[53] had a RMSE of 0.97 and 1.13 on the two test sets used in the study. This is accuracy is comparable to the one obtained in the present study with a minimal amount of datapoints. Han et al. used an end-to-end graph convolutional neural net[51] and managed to achieve a RMSE of 2.358 ppm for $^{13}$C-NMR chemical shifts on their test set, however, their study involved training set containing 32,609 of NMRShiftDB2 data points.

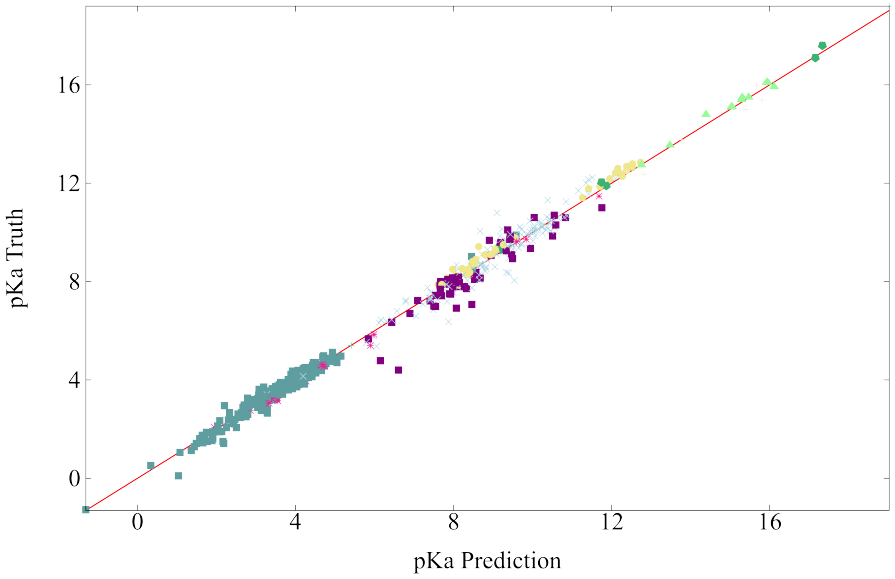
Qualitatively, the pK1/NMR values of the linear fit follow the expected chemical trend. For pK1, carboxylic acids groups populate the lower pKa scale and alcohols populate the higher end of the pKa scale. In middle part, there is a mix of carbamates, carbonates, and carbamides. These groups, particularly carbamides, are often part of arromatic rings in the IUPAC dataset thus lowering their pKa to a range that is more acidic than expected. The ability of the transfer model to predict this is evidence that the models holds enough long-range information (such as an entire aromatic groups) to be able to qualitatively differentiate when an acid is or is not part of an arromatic. The cutoff distance of 50 Å in the pre-trained model was important to include these non-local effects. In the linear NMR model, alkane groups populate the lower region of the predictions, whereas highly deshielded groups (those with oxygens especially) are shifted down the spectrum.

## Non-linear Transfer Learning – Embeddings to Atomistic pKa/NMR

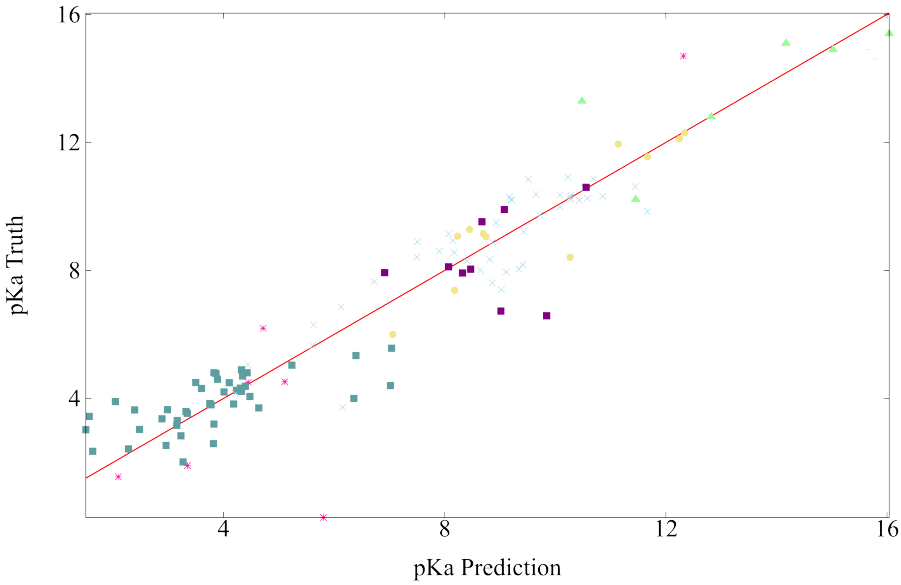
As a model with high intrinsic bias, linear regression is unlikely to provide a perfect model. In order to introduce more variance, we introduce non-linearity via a simple 2-layer feedforward neural net that maps each atomistic embedding to its local property. The training and testing results for both pKa and $^{13}$C-NMR can be seen in Figures 5 and 6. The train and test results gave a RMSE of 0.35 pKa units and 1.12 pKa units for pKa, respectively, and 6.01 ppm and 12.8 ppm for $^{13}$C-NMR, respectively. The results show much improvement after non-linearity is introduced.

## Size-Extensive Non-linear Transfer Learning – Embeddings to Solubility

Lastly, we tested the transferability of the entire molecular embedding representation to solubility. As described in the methods section, we designed a neural net architecture that computes size-extensive contributions from each embedding to a total log$S$ (logarithm of solubility measured in mol/L). Each embedding contributes an atomwise prediction that is summed to the total molecular solubility. The train and test results on logS predicted by our atomwise neural net (on a curated set of the NP-MRD database) is shown in Figure 7. The training RMSE is 0.02, whereas the testing RMSE is 0.67.

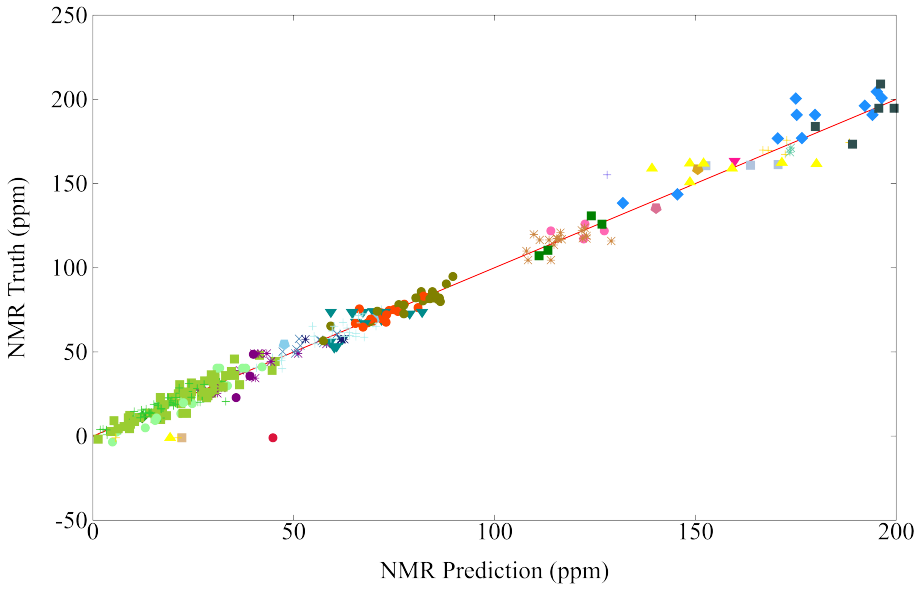



(a)

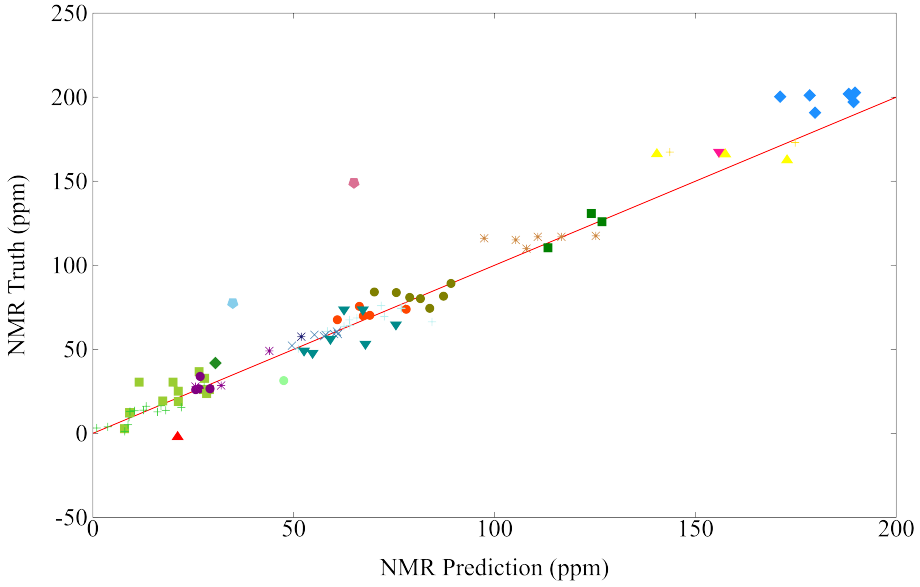



(b)

Figure 5: Predictions vs ground truth values for (a) training data (481 datapoints) and (b) test data (121 datapoints) from feed forward neural networks for pK1 from oxygen embeddings. Oxygen centered chemical environment labels are given in Figure 3c.

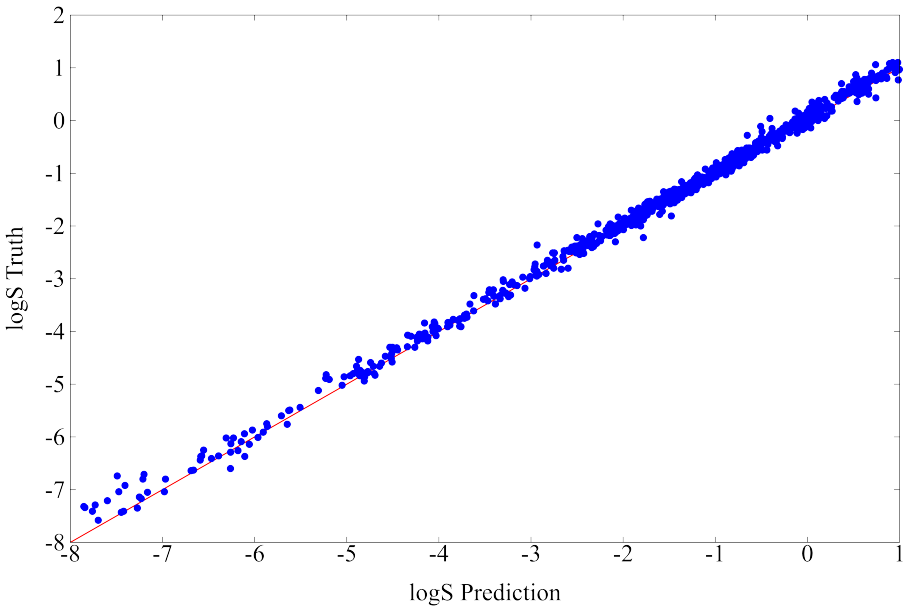



(a)

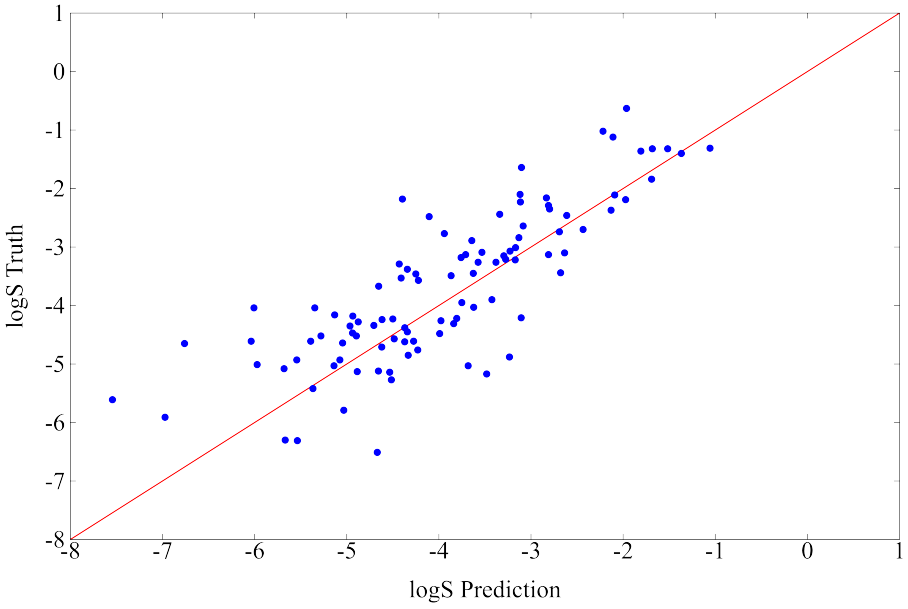



(b)

Figure 6: Predictions vs ground truth values for (a) training data (400 datapoints) and (b) test data (100 datapoints) from feed forward neural networks for $^{13}$C-NMR from carbon embeddings. Carbon centered chemical environment labels are given in Figure4c.




(a)




(b)

Figure 7: Predictions vs ground truth values for (a) training data (640 datapoints) and (b) test data (160 datapoints) from a size-extensive atomwize neural network for log$S$ solubility the entire molecule

The trained atomwise embeddings can then be visualized with Carbon-centred chemical environment labels to show an interpretable perspective of the model's decision-making in terms of each atom's contribution to the total molecular solubility. This can be done by first performing a principal component analysis (PCA)[96] on the embedding vectors which projects the data to a space that filters out the most important dimensions. In Figure 8, we plot log$S$ contribution from various atom-embeddings against the first principal dimension value of the embedding. This allows us to see how various atom-embeddings contribute differently to the log$S$ depending on which functional group they come from.

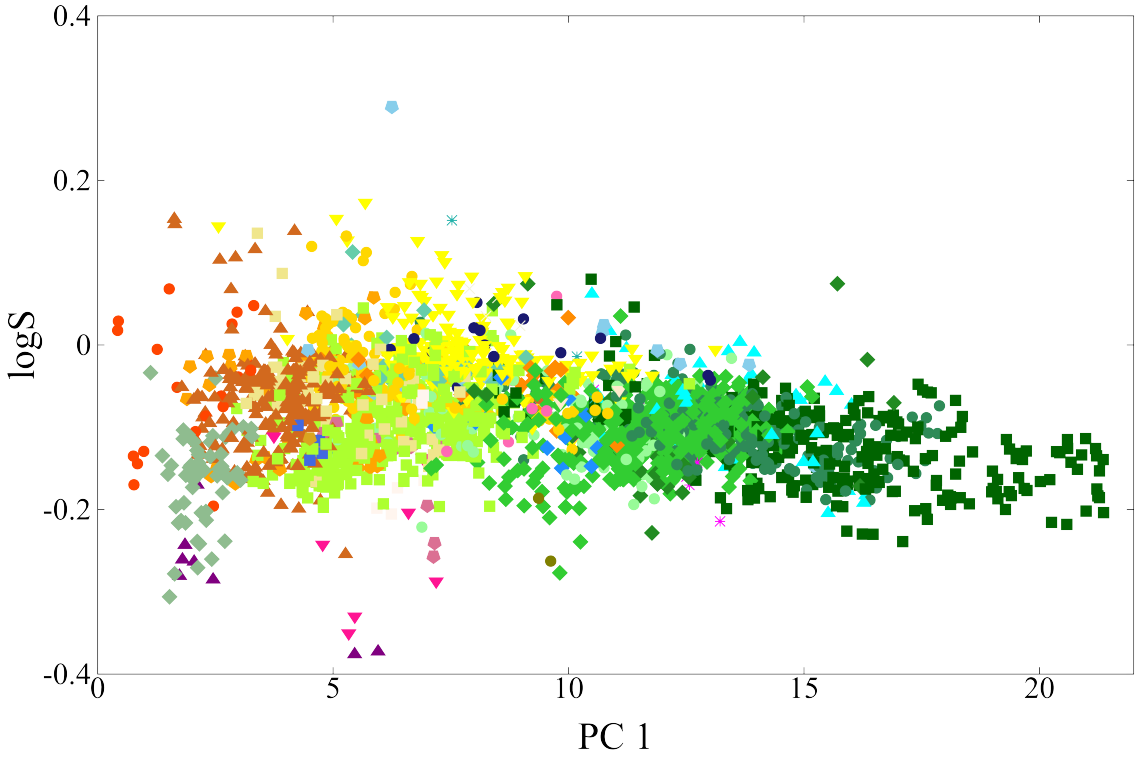
Some interesting trends can be noted from Figure 8. In general, as expected, groups with less symmetry contribute more to the total solubility prediction. For instance, it seems that carbons that have more hydrogens around them are generally more soluble (higher log$S$ value) than their unsaturated counterparts, possibly due to the asymmetry the hydrogen introduces. In addition, groups with only one hetero-atom around the carbon seem to be more soluble than groups with two/three hetero-atoms around the carbon. This could be attributed to in enhanced polarizability of moieties with a single hetero-atom, whereas the polarity can be neutralized with additional hetero-atoms, consequently bringing the solubility down. This is most evident by the placement of tri-amine substituted carbon in Figure 8, which is apparently the least soluble group according to the model.

Notably, the Graph-pKa model by Xiong et al.[54] which predicts macro-pKa from the entire molecular representation was also able to automatically deconvolute the macro-pKa into discrete micro-pKa values by visualization of the atom-embeddings through a 3D PCA projection of the functional-group-labeled atom embeddings. This effect, where the target molecular representation and a proposed atomistic representation are simultaneously learned, has been noted in previous works but never extensively studied as a stand-alone property of graph-based molecular representations,[97;98] especially since it interfaces naturally for atomistic systems. For example, Schütt et al.[97] used a probe atom that acted as a test charge. This probe was used to analyze the SchNet graph convolutional model in 3D space. From this, they were able to provide a spatial heat map of the contributions to the potential energy for every point around the molecule. This effect is not restricted to graph-based models as, for example, Rasmussen et al.[98] found that they could also obtain atomistic contributions from a random forest model trained on molecular descriptors for the log$P$ solubility. The atomistic model that is being probed in these works is built by the graph model during training and is capturing correlations across the atoms in the molecule.

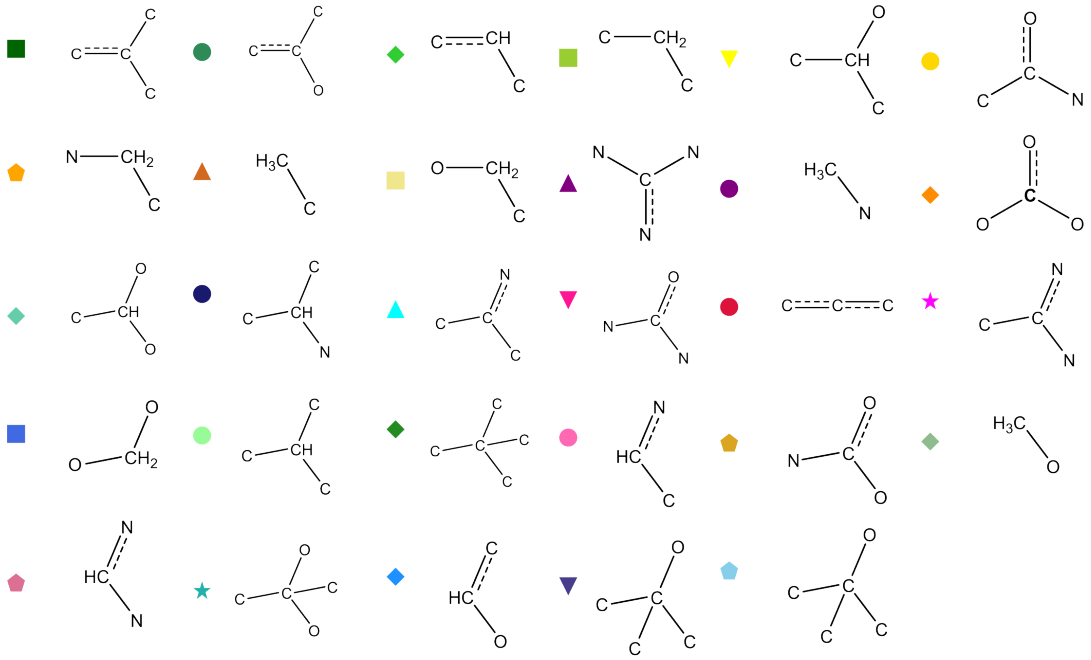
# Conclusions

Transferable learning is a promising direction for GCNN-based machine learning models in chemistry as it provides an opportunity to build and evaluate a possibly unified and complete (statistical) representation of chemistry, one that can be used for a diverse set of tasks.

From previous work,[81] we saw that embeddings recognize chemical environments with high accuracy




(a) log$S$ vs PC 1 of carbon-embedding data.



(b) Carbon-centric functional groups found in the curated NP-MRD database.

Figure 8: (a) The log$S$ contribution of each carbon-embedding plotted against the first principal component of the C-embedding, labeled according to C-centered functional group (b)

and can give a measure of chemical and structural similarity between molecules. In this work, we expand on this by showing that such a representation can generalize to predict a range of atomistic properties and molecular properties using an intermediatte transfer learning approach that takes embeddings as atomistic descriptors and predicts the associated atomistic/molecular property.

We tested the transferability of graph-built atomistic embedding representation to predict a diverse set of chemical properties: pK1, $^{13}$C-NMR, and log$S$. The results show some promise that pre-trained neural network representations can indeed be used to model a larger range of chemical observables. Whereas atomistic properties, such as pKa and $^{13}$C-NMR, the transfer models can be built directly on the atomistic embeddings, molecular properties require the design of size-extensive transfer models to predict contributions from each atom towards the total target molecular property.

In general while our results do not reach perfect accuracy, the datasets used in the study are very small, which is a challenge for neural networks in general and graph-based ones in particular. Regardless of this challenge, our transfer models find it easy to learn qualitative chemical information from energy-trained embeddings on small datasets. Future work will expand on this by curating larger databases and further finetuning the transfer models.

The search for a global statistical representation is important as it would standardize much of the plethora of neural network representations (and neural network models) for chemistry and may allow chemists to achieve a unified baseline (and generaly understood) model. The embedding representation is a clear candidate for this search, as it is conceptually simple as an atom-based model, while also allowing for size-extensive extensions to model a wider range of molecular properties whilst remaining interpretable to the user of the model.

## Competing interests statement

The authors declare no competing interests.

## Author contribution statement

SDC designed and investigated transfer learning models for pKa, BM designed and investigated transfer learning models for NMR. AES designed and transfer learning models for log$S$. AES provided guidance and mentoring to undergraduate students SDC and BM. SDB supervised the project. AES wrote the first draft of the manuscript, whereas all authors co-wrote and edited the manuscript.

## Acknowledgements

BM acknowledges support from Stefan Kuhn in accessing the NMRShiftDB2 database.

## Funding statement

SDB and AES acknowledge the Canada Research Chair program, the CFI, NSERC, and NBIF for financial suppport.

## Data availability statement

For embedding vector data of QM9 molecules, and the trained SchNet model used, see: https://doi.org/10.25545/EK1EQA. [85]

Neural net codes can be found at: https://github.com/amerelsamman/Transfer-Learning-Graph-Representations-of-Molecules-for-pKa-13C-NMR-and-Solubility.

Other data generated or analyzed during this study (such as the curated datasets) are available from the corresponding author upon reasonable request.

## References

[1] Kristof T Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8:1, 2017.

[2] Ryo Fujiki, Toru Matsui, Yasuteru Shigeta, Haruyuki Nakano, and Norio Yoshida. Recent developments of computational methods for pka prediction based on electronic structure theory with solvation models. *Multi-disciplinary Scientific Journal*, 4:849, 2021.

[3] Felipe Ribeiro Dutra and Rogério Custodio. Development of a direct method to calculate pka using electronic structure methods.

[4] Junming Ho, Vincent E. Zwicker, Karen K. Y. Yuen, and Katrina A. Jolliffe. Quantum chemical prediction of equilibrium acidities of ureas, deltamides, squaramides, and croconamides. *Journal of Organic Chemistry*, 82:10732, 2017.

[5] Paul Popelier. pka prediction from ab initio calculations. *Research Outreach*, 2019.

[6] Nguyen Thi My, Nguyen Van Din, and Mai Van Bay. The prediction of pka values for phenolic compounds by the dft theory. *UD-JST*, 20:50, 2022.

[7] Samarjeet Prasad, Jing Huang, Qiao Zeng, and Bernard R. Brooks. An explicit-solvent hybrid qm and mm approach for predicting pka of small molecules in sampl6 challenge. *Journal of Computer-Aided Molecular Design*, 32:1191, 2018.

[8] Raymond J Abraham and Mehdi Mobli. *Modelling 1H NMR spectra of organic compounds: theory, applications and NMR prediction software*. John Wiley & Sons, 2008.

[9] Eugene E Kwan and Richard Y Liu. Enhancing nmr prediction for organic compounds using molecular dynamics. *Journal of Chemical Theory and Computation*, 11:5083, 2015.

[10] Jens J Led and Henrik Gesmar. Application of the linear prediction method to nmr spectroscopy. *Chemical Reviews*, 91:1413, 1991.

[11] P Koehl. Linear prediction spectral analysis of nmr data. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 34:257, 1999.

[12] J Tomasi, Benedetta Mennucci, and E Cancès. The ief version of the pcm solvation method: an overview of a new method addressed to study molecular solutes at the qm ab initio level. *Journal of Molecular Structure: THEOCHEM*, 464:211, 1999.

[13] David J Tannor, Bryan Marten, Robert Murphy, Richard A Friesner, Doree Sitkoff, Anthony Nicholls, Barry Honig, Murco Ringnalda, and William A Goddard III. Accurate first principles calculation of molecular charge distributions and solvation energies from ab initio quantum mechanics and continuum dielectric theory. *Journal of the American Chemical Society*, 116:11875, 1994.

[14] Lavanya M Ramaniah, Marco Bernasconi, and Michele Parrinello. Ab initio molecular-dynamics simulation of k+ solvation in water. *The Journal of Chemical Physics*, 111:1587, 1999.

[15] Jimmy C Kromann, Frej Larsen, Hadeel Moustafa, and Jan H Jensen. Prediction of pka values using the pm6 semiempirical method. *PeerJ*, 4:2335, 2016.

[16] Cornelis Matijssen. A comparative study of in silico prediction of pka. *Journal of Cheminformatics*, 2:37, 2010.

[17] Douglas Dalzell Perrin, Boyd Dempsey, and Eldon Perey Serjeant. *pKa prediction for organic acids and bases*, volume 1. Springer, 1981.

[18] Elmar Krieger, Jens E Nielsen, Chris AEM Spronk, and Gert Vriend. Fast empirical pka prediction by ewald summation. *Journal of Molecular Graphics and Modelling*, 25:481, 2006.

[19] George C Shields and Paul G Seybold. *Computational approaches for the prediction of pKa values*. CRC Press, 2013.

[20] Tomasz Puzyn, Aleksandra Mostrag, Jerzy Falandysz, Yana Kholod, and Jerzy Leszczynski. Predicting water solubility of congeners: chloronaphthalenes—a case study. *Journal of Hazardous Materials*, 170:1014, 2009.

[21] William L Jorgensen and Erin M Duffy. Prediction of drug solubility from monte carlo simulations. *Bioorganic & Medicinal Chemistry Letters*, 10:1155, 2000.

[22] Jogoth Ali, Patrick Camilleri, Marc B Brown, Andrew J Hutt, and Stewart B Kirton. Revisiting the general solubility equation: in silico prediction of aqueous solubility incorporating the effect of topographical polar surface area. *Journal of Chemical Information and Modeling*, 52:420, 2012.

[23] Eric Jonas, Stefan Kuhn, and Nils Schlörer. Prediction of chemical shift in nmr: A review. *Magnetic Resonance in Chemistry*, 60:1021, 2022.

[24] Carlos Cobas. Nmr signal processing, prediction, and structure verification with machine learning techniques. *Magnetic Resonance in Chemistry*, 58:512, 2020.

[25] Stefan Kuhn, Björn Egert, Steffen Neumann, and Christoph Steinbeck. Building blocks for automated elucidation of metabolites: Machine learning methods for nmr prediction. *BMC Bioinformatics*, 9:1, 2008.

[26] João Aires-de Sousa, Markus C Hemmer, and Johann Gasteiger. Prediction of 1h nmr chemical shifts using neural networks. *Analytical Chemistry*, 74:80, 2002.

[27] Will Gerrard, Lars A Bratholm, Martin J Packer, Adrian J Mulholland, David R Glowacki, and Craig P Butts. Impression–prediction of nmr parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. *Chemical Science*, 11:508, 2020.

[28] Yang Shen and Ad Bax. Sparta+: a modest improvement in empirical nmr chemical shift prediction by means of an artificial neural network. *Journal of Biomolecular NMR*, 48:13, 2010.

[29] Robert Fraczkiewicz, Mario Lobell, Andreas H. Göller, Ursula Krenz, Rolf Schoenneis, Robert D. Clark, and Alexander Hillisch. Best of both worlds: Combining pharma data and state of the art modeling technology to improve in silico pka prediction. *Journal of Chemical Information and Modeling*, 55:389, 2015.

[30] Mengshan Li, Huaijing Zhang, Bingsheng Chen, Yan Wu, and Lixin Guan. Prediction of pka values for neutral and basic drugs based on hybrid artificial intelligence methods. *Scientific Reports*, 8, 2018.

[31] Jialu Wu, Yu Kang, Peichen Pan, and Tingjun Hou. Machine learning methods for pka prediction of small molecules: Advances and challenges. *Drug Discovery Today*, page 103372, 2022.

[32] Samuel Boobier, David RJ Hose, A John Blacker, and Bao N Nguyen. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nature Communications*, 11:5753, 2020.

[33] Hongmao Sun. A universal molecular descriptor system for prediction of logp, logs, logbb, and absorption. *Journal of Chemical Information and Computer Sciences*, 44:748, 2004.

[34] Emad A. S. Al-Hyali, Nezar A. Al-Azzawi, and Faiz M. Al-Abady. Statistical study for the prediction of pka values of substituted benzaldoxime based on quantum chemicals methods. *Journal of the Korean Chemical Society*, 55:733, 2011.

[35] Qi Yang, Yao Li, Jin-Dong Yang, Yidi Liu, Long Zhang, Sanzhong Luo, and Jin-Pei Cheng. Holistic prediction of the pka in diverse solvents based on a machine-learning approach. *Angewandte Chemie*, 132:19444, 2020.

[36] Jesús Jover, Ramón Bosque, and Joaquim Sales. Qspr prediction of pka for benzoic acids in different solvents. *QSAR & Combinatorial Science*, 27:563, 2008.

[37] Jesús Jover, Ramón Bosque, and Joaquim Sales. Neural network based qspr study for predicting pka of phenols in different solvents. *QSAR & Combinatorial Science*, 26:385, 2007.

[38] João Aires-de Sousa, Markus C Hemmer, and Johann Gasteiger. Prediction of 1h nmr chemical shifts using neural networks. *Analytical Chemistry*, 74:80, 2002.

[39] Robert C Glen, Andreas Bender, Catrin H Arnby, Lars Carlsson, Scott Boyer, and James Smith. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to adme. *IDrugs*, 9:199, 2006.

[40] Anna Vulpetti, Gregory Landrum, Simon Rüdisser, Paulus Erbel, and Claudio Dalvit. 19f nmr chemical shift prediction with fluorine fingerprint descriptor. *Journal of Fluorine Chemistry*, 131:570, 2010.

[41] Jesús Jover, Ramón Bosque, and Joaquim Sales. Neural network based qspr study for predicting pka of phenols in different solvents. *QSAR & Combinatorial Science*, 26:385, 2007.

[42] Han van de Waterbeemd, Gian Camenisch, Gerd Folkers, Jacques R Chretien, and Oleg A Raevsky. Estimation of blood-brain barrier crossing of drugs using molecular size and shape, and h-bonding descriptors. *Journal of Drug Targeting*, 6:151, 1998.

[43] Dejun Jiang, Zhenxing Wu, Chang-Yu Hsieh, Guangyong Chen, Ben Liao, Zhe Wang, Chao Shen, Dongsheng Cao, Jian Wu, and Tingjun Hou. Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. *Journal of Cheminformatics*, 13:1, 2021.

[44] Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Marvin Steijaert, Jörg K Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chemical Science*, 9:5441, 2018.

[45] Bowen Tang, Skyler T Kramer, Meijuan Fang, Yingkun Qiu, Zhen Wu, and Dong Xu. A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility. *Journal of Cheminformatics*, 12:1, 2020.

[46] Vadim Korolev, Artem Mitrofanov, Alexandru Korotcov, and Valery Tkachenko. Graph convolutional neural networks as "general-purpose" property predictors: the universality and limits of applicability. *Journal of Chemical Information and Modeling*, 60:22, 2019.

[47] Junshui Ma, Robert P Sheridan, Andy Liaw, George E Dahl, and Vladimir Svetnik. Deep neural nets as a method for quantitative structure–activity relationships. *Journal of Chemical Information and Modeling*, 55:263, 2015.

[48] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59:3370, 2019.

[49] Junying Li, Deng Cai, and Xiaofei He. Learning graph-level representation for drug discovery. *arXiv preprint arXiv:1709.03741*, 2017.

[50] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9:513.

[51] Jongmin Han, Hyungu Kang, Seokho Kang, Youngchun Kwon, Dongseon Lee, and Youn-Suk Choi. Scalable graph neural network for nmr chemical shift prediction. *Physical Chemistry Chemical Physics*, 24:26870, 2022.

[52] Youngchun Kwon, Dongseon Lee, Youn-Suk Choi, Myeonginn Kang, and Seokho Kang. Neural message passing for nmr chemical shift prediction. *Journal of Chemical Information and Modeling*, 60:2024, 2020.

[53] Fritz Mayr, Marcus Wieder, Oliver Wieder, and Thierry Langer. Improving small molecule pka prediction using transfer learning with graph neural networks. *Frontiers in Chemistry*, 10:866585, 2022.

[54] Jiacheng Xiong, Zhaojun Li, Guangchao Wang, Zunyun Fu, Feisheng Zhong, Tingyang Xu, Xiaomeng Liu, Ziming Huang, Xiaohong Liu, Kaixian Chen, et al. Multi-instance learning of graph neural networks for aqueous p k a prediction. *Bioinformatics*, 38:792, 2022.

[55] Dongdong Zhang, Song Xia, and Yingkai Zhang. Accurate prediction of aqueous free solvation energies using 3d atomic feature-based graph neural network with transfer learning. *Journal of Chemical Information and Modeling*, 62:1840, 2022.

[56] Yashaswi Pathak, Sarvesh Mehta, and U Deva Priyakumar. Learning atomic interactions through solvation free energy prediction using graph neural networks. *Journal of Chemical Information and Modeling*, 61:689, 2021.

[57] Oliver Wieder, Stefan Kohlbacher, Mélaine Kuenemann, Arthur Garon, Pierre Ducrot, Thomas Seidel, and Thierry Langer. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*, 37:1, 2020.

[58] Kaycee Low, Michelle L Coote, and Ekaterina I Izgorodina. Explainable solvation free energy prediction combining graph neural networks with chemical intuition. *Journal of Chemical Information and Modeling*, 62:5457, 2022.

[59] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, page 1263, 2017.

[60] Youngchun Kwon, Dongseon Lee, Youn-Suk Choi, Myeonginn Kang, and Seokho Kang. Neural message passing for nmr chemical shift prediction. *Journal of Chemical Information and Modeling*, 60:2024, 2020.

[61] Laurianne David, Amol Thakkar, Rocío Mercado, and Ola Engkvist. Molecular representations in ai-driven drug discovery: a review and practical guide. *Journal of Cheminformatics*, 12:1, 2020.

[62] Oliver T Unke and Markus Meuwly. Physnet: A neural network for predicting energies, forces, dipole moments, and partial charges. *Journal of Chemical Theory and Computation*, 15:3678, 2019.

[63] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57, 2020.

[64] Oliver Wieder, Stefan Kohlbacher, Mélaine Kuenemann, Arthur Garon, Pierre Ducrot, Thomas Seidel, and Thierry Langer. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*, 37:1, 2020.

[65] Yuyang Wang, Zijie Li, and Amir Barati Farimani. Graph neural networks for molecules. *arXiv preprint arXiv:2209.05582*, 2022.

[66] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in Neural Information Processing Systems*, 30, 2017.

[67] KT Schutt, Pan Kessel, Michael Gastegger, KA Nicoli, Alexandre Tkatchenko, and K-R Muüller. Schnetpack: A deep learning toolbox for atomistic systems. *Journal of Chemical Theory and Computation*, 15:448, 2018.

[68] Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148:241722, 2018.

[69] Yanfei Guan, SV Shree Sowndarya, Liliana C Gallegos, Peter C St John, and Robert S Paton. Real-time prediction of 1 h and 13 c chemical shifts with dft accuracy using a 3d graph neural network. *Chemical Science*, 12:12012, 2021.

[70] Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems*, 33:7852, 2020.

[71] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109:43, 2020.

[72] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109:43, 2020.

[73] Hironao Yamada, Chang Liu, Stephen Wu, Yukinori Koyama, Shenghong Ju, Junichiro Shiomi, Junko Morikawa, and Ryo Yoshida. Predicting materials properties with little data using shotgun transfer learning. *ACS Central Science*, 5:1717, 2019.

[74] Chenjing Cai, Shiwei Wang, Youjun Xu, Weilin Zhang, Ke Tang, Qi Ouyang, Luhua Lai, and Jianfeng Pei. Transfer learning for drug discovery. *Journal of Medicinal Chemistry*, 63:8683, 2020.

[75] Rahul Barman, Sharvari Deshpande, Shruti Agarwal, Unzela Inamdar, M Devare, and A Patil. Transfer learning for small dataset. In *Proceedings of the National Conference on Machine Learning*, volume 26. ResearchGate Berlin, Germany, 2019.

[76] Colin A Grambow, Yi-Pei Li, and William H Green. Accurate thermochemistry with small data sets: A bond additivity correction and transfer learning approach. *The Journal of Physical Chemistry A*, 123:5826, 2019.

[77] Joshua L Lansford, Brian C Barnes, Betsy M Rice, and Klavs F Jensen. Building chemical property models for energetic materials from small datasets using a transfer learning approach. *Journal of Chemical Information and Modeling*, 62:5397, 2022.

[78] Justin S Smith, Benjamin T Nebgen, Roman Zubatyuk, Nicholas Lubbers, Christian Devereux, Kipton Barros, Sergei Tretiak, Olexandr Isayev, and Adrian Roitberg. Outsmarting quantum chemistry through transfer learning. *ChemRxiv preprint 10.26434/chemrxiv.6744440.v1*, 2018.

[79] Herim Han and Sunghwan Choi. Transfer learning from simulation to experimental data: Nmr chemical shift predictions. *The Journal of Physical Chemistry Letters*, 12:3662, 2021.

[80] Florence H Vermeire and William H Green. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chemical Engineering Journal*, 418:129307, 2021.

[81] Amer El Samman, Incé Amina Husain, Mai Huynh, Stefano De Castro, Brooke Morton, Guillaume Acke, and Stijn De Baerdemacker. Global interpretability and geometry of graph convolutional neural networks for chemistry in terms of chemical moieties. *chemrxiv*, 2023.

[82] Thomas Mensink, Jasper Uijlings, Alina Kuznetsova, Michael Gygli, and Vittorio Ferrari. Factors of influence for transfer learning across diverse appearance domains and task types. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:9298, 2021.

[83] Daniel Bolya, Rohit Mittapalli, and Judy Hoffman. Scalable diverse model selection for accessible transfer learning. *Advances in Neural Information Processing Systems*, 34:19301, 2021.

[84] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1:1, 2014.

[85] Amer Marwan El-Samman. SchNet Model Embedding Vectors of QM9 Atoms Labelled According to Functional Groups Designation, 2023.

[86] Jonathan Zheng. Iupac/dissociation-constants: v1.0, 2022.

[87] Douglas Dalzell Perrin. *Dissociation Constants of Organic Bases in Aqueous Solution: Hauptbd.* Butterworths, 1965.

[88] Douglas Dalzell Perrin. *Dissociation constants of organic bases in aqueous solution*, volume 1. Pergamon, 1972.

[89] Douglas Dalzell Perrin. *Ionisation Constants of Organic Acids in Aqueous Solution*, volume 1. Pergamon, 1972.

[90] Stefan Kuhn and Nils E Schlörer. Facilitating quality control for spectra assignments of small organic molecules: nmrshiftdb2–a free in-house nmr database with integrated lims for academic service laboratories. *Magnetic Resonance in Chemistry*, 53:582, 2015.

[91] Christoph Steinbeck, Stefan Krause, and Stefan Kuhn. Nmrshiftdb constructing a free chemical information system with open-source components. *Journal of Chemical Information and Computer Sciences*, 43:1733, 2003.

[92] Christoph Steinbeck and Stefan Kuhn. Nmrshiftdb–compound identification and structure elucidation support through a free community-built web database. *Phytochemistry*, 65:2711, 2004.

[93] David S Wishart, Zinat Sayeeda, Zachary Budinski, AnChi Guo, Brian L Lee, Mark Berjanskii, Manoj Rout, Harrison Peters, Raynard Dizon, Robert Mah, et al. Np-mrd: the natural products magnetic resonance database. *Nucleic Acids Research*, 50:665, 2022.

[94] W Bremser. Hose—a novel substructure code. *Analytica Chimica Acta*, 103:355, 1978.

[95] Anna Gaulton, Anne Hersey, Michał Nowotka, A Patricia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J Bellis, Elena Cibrián-Uhalte, et al. The chembl database in 2017. *Nucleic Acids Research*, 45:945, 2017.

[96] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:433, 2010.

[97] Kristof T Schütt, Alexandre Tkatchenko, and Klaus-Robert Müller. Learning representations of molecules and materials with atomistic neural networks. *Machine Learning Meets Quantum Physics*, page 215, 2020.

[98] Maria H Rasmussen, Diana S Christensen, and Jan H Jensen. Do machines dream of atoms? crippen's logp as a quantitative molecular benchmark for explainable ai heatmaps. *SciPost Chemistry*, 2:2, 2023.