liquids

MDPI

Article

AbraLlama: Predicting Abraham Model Solute Descriptors and Modified Solvent Parameters Using Llama

Andrew S. I. D. Lang * and Youngmin Lee

Department of Computing & Mathematics, Oral Roberts University, Tulsa, OK 74171, USA; ttmn_ym@oru.edu
* Correspondence: alang@oru.edu

Abstract: This study explores the application of fine-tuned large language models for predicting physicochemical properties, specifically focusing on Abraham model solute descriptors (E, S, A, B, V) and modified solvent parameters (e_0, s_0, a_0, b_0, v_0). By leveraging ChemLLaMA, a specialized version of the LLaMA model for cheminformatics tasks, we developed the AbraLlama-Solvent and AbraLlama-Solute models using curated datasets of experimentally derived solute descriptors and solvent parameters. Our findings demonstrate that AbraLlama-Solvent and AbraLlama-Solute predict modified solvent parameters and solute descriptors with high accuracy, comparable to existing methods. The AbraLlama-Solvent model shows varying prediction accuracy across different solvents, influenced by their position within the chemical space, while the AbraLlama-Solute model consistently predicts solute descriptors with high accuracy. Both models are available as applications on Hugging Face, facilitating easy predictions from SMILES strings. This research highlights the potential of LLMs in chemistry applications, offering practical tools for solvent comparison and expanding the applicability of Abraham solvation equations to a broader range of organic solvents.

Keywords: Abraham general solvation model; Abraham solute descriptors; Abraham solvent parameters; solvent comparison; solvent replacement; ChemLLaMA

check for updates

Citation: Lang, A.S.I.D.; Lee, Y. AbraLlama: Predicting Abraham Model Solute Descriptors and Modified Solvent Parameters Using Llama. *Liquids* **2024**, *4*, 518–524. https://doi.org/10.3390/liquids4030029

Academic Editors: William E. Acree, Jr., Juan Ortega Saavedra and Enrico Bodo

Received: 24 May 2024
Revised: 19 June 2024
Accepted: 29 July 2024
Published: 2 August 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The use of Artificial Intelligence (AI), including using fine-tuned large language models (LLMs), for predicting molecular properties has garnered significant attention in recent years [1,2]. LLMs are advanced AI systems that utilize deep learning techniques, specifically transformer architectures, to process and understand large amounts of textual data. These models are initially pre-trained on extensive corpora to capture general language patterns and semantic relationships, which are then fine-tuned on domain-specific datasets for specialized applications, such as cheminformatics; see the recent article by Luong and Singh for a review of how transformer models are being used in cheminformatics [3]. Building on this interest, our study leverages the capabilities of ChemLLaMA [4], a specialized version of the LLaMA (Large Language Model Meta AI) tailored for cheminformatics tasks, to predict Abraham model solute descriptors (E, S, A, B, V) and modified Abraham solvent parameters (e_0, s_0, a_0, b_0, v_0) [5]. This aligns with recent advancements in machine learning models for predicting solute parameters and solvation energies, which have shown promising results in both accuracy and applicability, such as the SoluteGC, SoluteML, and DirectML models [6].

The Abraham general solvation model is integral to predicting solubility and partition coefficients across a wide range of solvation systems. The model is based on linear free energy relationships (LFERs) between solute descriptors and solvent parameters [7–10]. This model is expressed as

$$\log P = c + e \cdot E + s \cdot S + a \cdot A + b \cdot B + v \cdot V, \quad (1)$$

where log P represents the solvent/water partition coefficient. Under typical conditions, this model can also predict the solubility of organic compounds in organic solvents [9] as follows:

$$\log S_s = \log S_w + c + e \cdot E + s \cdot S + a \cdot A + b \cdot B + v \cdot V, \quad (2)$$

where S_s is the molar concentration of the solute in the organic solvent, and S_w is the molar concentration of the solute in water. The parameters c, e, s, a, and b are the standard Abraham solvent parameters, and E, S, A, B, and V are the Abraham solute descriptors: E is the solute excess molar refractivity in units of $(cm^3/mol)/10$, S is the solute dipolarity/polarizability, A and B are the overall or summation hydrogen bond acidity and basicity, and V is the McGowan characteristic volume in units of $(cm^3/mol)/100$.

To determine the Abraham solvent parameters, linear regression is typically performed on the partition coefficients and solubilities of solutes with known Abraham descriptors, which have been experimentally measured in the solvent. The intercept c (see Equation (1)) is traditionally allowed to float, ideally capturing information not characterized by the other solvent–solute interaction terms. However, this approach complicates direct comparisons of different solvents. To address this, we utilized modified Abraham solvent parameters (e_0, s_0, a_0, b_0, v_0), derived by regressing with the intercept set to zero, as described by Bradley et al. [5], to facilitate more straightforward comparisons.

Identifying alternative solvents for various processes is essential due to the significant environmental and health risks associated with some traditional solvents. Increased regulatory demands; the push for sustainability; reactivity; volatility; and cost considerations are also critical factors in solvent selection. While several solvent replacement tools and lists exist [11], this work aims to provide an open tool for comparing solvents based on their modified Abraham solvent parameters. Solvents with closely matching parameters are likely to exhibit similar solvation properties [5].

Thus, the objectives of this study are twofold: firstly, to demonstrate how large language models (LLMs) can predict physicochemical properties, specifically, modified Abraham solvent parameters and Abraham solute descriptors, and secondly, to provide open models (AbraLlama-Solvent and AbraLlama-Solute) that facilitate the prediction of these parameters and descriptors. This extends the applicability of the Abraham solvation equations to a broader range of organic solvents and aids in identifying solvents with comparable solvation properties.

Our findings demonstrate that the models predict modified Abraham solvent parameters and Abraham solute descriptors with accuracy comparable to existing methods. Furthermore, by making these models available as apps on Hugging Face, a platform that provides open-source tools and models for the easy deployment and sharing of machine learning models, we provide a valuable service to the community, enabling easy and accessible calculations without requiring users to have expertise in AI model deployment [12]. This research highlights the significant potential of LLMs in chemistry applications and offers a practical tool for solvent comparison.

2. Materials and Methods

2.1. Datasets

Experimentally determined Abraham solute descriptors (E, S, A, B, V) were obtained from the UFZ-LSER database (version 3.2.1) using the search parameter "%C%" and with the radio button "experimental descriptors" selected [13]. Standard Abraham solvent parameters (c, e, s, a, b, v) were provided by Dr. William E. Acree Jr., who compiled them from the literature. We have made this solvent dataset, which contains both Abraham model log P and log K (not used in this work) equation coefficients for both pure solvents and mixtures, available under a CC0 license on Figshare [14], where the subscripts 3 and 4 in the dataset refer to equations involving log P and log K respectively [15].

2.2. Data Preprocessing

The Abraham solute descriptor dataset contained multiple records for many compounds. To ensure consistency, we filtered the data and kept only entries labeled as "LSER Dataset for CompTox users (2017)". This resulted in a final dataset of N = 6852 compounds with experimentally derived Abraham solute descriptors. We excluded mixtures from the Abraham solvent parameter dataset and kept only pure solvents (N = 122) with experimentally derived solvent parameters. Since the c-parameter can complicate direct comparisons between solvents, we calculated modified Abraham solvent parameters (e_0, s_0, a_0, b_0, v_0) for all solvents using the method described by Bradley et al. [5]. This process involved calculating log P values for all solutes in each solvent using Equation (1). We then determined the modified Abraham solvent parameters by regressing these log P values with a linear equation with zero intercept: $\log P = e_0 \cdot E + s_0 \cdot S + a_0 \cdot A + b_0 \cdot B + v_0 \cdot V$. Both the original and modified Abraham solvent parameters for the modeling dataset are available under a CC0 license from Figshare [16].

2.3. Model Development

We leveraged ChemLLaMA, a fine-tuned version of the LLaMA transformer model, which was developed to predict molecular properties from SMILES strings [4]. This study utilized the 30-million-parameter ChemLLaMA model [17] and further fine-tuned it to predict both Abraham solute descriptors and modified Abraham solvent parameters. Note that we did not freeze the pre-trained model for fine-tuning for this research. We refer the reader to the literature for a more detailed description of ChemLLaMA, including the source code [4,17].

To create the AbraLlama-Solvent and AbraLlama-Solute models presented here, we fine-tuned ChemLLaMA using the curated datasets mentioned above. Training occurred on a single GPU (NVIDIA A30) using PyTorch, specifically PyTorch-Lightning (v. 1.9.5) [18] and Lightning-Bolts (v. 0.7.0) [19]. Both models underwent 20 epochs of training with a learning rate of 0.0001 with a 'Linear Warmup Cosine Annealing Learning Rate' scheduler which reaches the peak learning rate at the second epoch. The training setup for the AbraLlama-Solute model used 5-fold cross-validation and a batch size pair of (64, 64). Due to the smaller dataset size, 10-fold cross-validation and a batch size pair of (10, 10) were employed for training the AbraLlama-Solvent model. Cross-validation allowed us to report test-set-equivalent statistics.

Both models were retrained on the complete, unpartitioned datasets to maximize predictive accuracy and overall utility. These AbraLlama models are available as applications on Hugging Face, allowing users to predict modified solvent parameters and solute descriptors from SMILES string inputs [12].

Our methodology involved pre-training transformer models on extensive datasets, followed by fine-tuning them on task-specific datasets, ensuring consistent training environments for reliable comparisons. The key steps included using uniform tokenizers, standardized training data, and identical training configurations across models. Advanced optimization techniques and learning rate schedulers were employed to ensure effective model convergence. Additionally, rigorous cross-validation and repeated fine-tuning iterations were used to mitigate statistical anomalies and enhance robustness. This structured approach, which encompasses dataset preparation, model pre-training, fine-tuning, and thorough evaluation, provides a replicable framework adaptable for model comparison and fine-tuning in various domains. For more details, refer to the ChemLLaMA paper [4] and the Python source code for ChemLLaMA [17] and AbraLlama-Solvent and AbraLlama-Solute available on GitHub [20].

3. Results and Discussion

The ChemLLaMA model [4], originally fine-tuned from the LLaMA large language model (LLM) for cheminformatics applications, was further refined to develop models capable of predicting Abraham model solute descriptors (E, S, A, B, V) and modified Abraham

solvent parameters (e_0, s_0, a_0, b_0, v_0). The performance of these fine-tuned ChemLLaMA models, designated as AbraLlama-Solvent and AbraLlama-Solute, was evaluated using cross-validation. Table 1 presents the aggregated cross-validated statistics from all folds, demonstrating the models' ability to predict new molecules with accuracy comparable to other established methods [5,13].

Table 1. Cross-validated statistics for the AbraLlama-Solvent and AbraLlama-Solute models.

Modified Solvent Parameters				Solute Descriptors			
Parameter	***N***	**RMSE**	**R^2**	**Descriptor**	***N***	**RMSE**	**R^2**
e_0	122	0.163	0.32	E	6852	0.132	0.97
s_0	122	0.353	0.66	S	6852	0.240	0.90
a_0	122	0.655	0.81	A	6852	0.135	0.85
b_0	122	0.480	0.40	B	6852	0.123	0.96
v_0	122	0.318	0.49	V	6852	0.097	0.98

3.1. AbraLlama-Solvent

The AbraLlama-Solvent model predicts modified Abraham solvent parameters (e_0, s_0, a_0, b_0, v_0) with varying degrees of accuracy depending on the solvent's position within the chemical space, as determined using principal component analysis (PCA). The region around the origin—the center of the chemical space—is populated by acyclic alcohols, esters, and ethers consisting solely of carbon, oxygen, and hydrogen atoms (C, O, H). In contrast, the outer regions include compounds with rings and those containing nitrogen, sulfur, and fluorine (N, S, F).

The prediction error (the Euclidean distance from the predicted to measured values) is correlated with the distance from the center of the chemical space ($p < 0.0001$), as shown in Figure 1. Four clear outliers—carbon disulfide, trifluoroethanol, pyridine, and dimethyl sulfoxide—exhibit significant deviations from the model's predictions. When these outliers are excluded, the model's cross-validated statistics improve markedly: e_0 (RMSE = 0.154, R^2 = 0.364), s_0 (RMSE = 0.330, R^2 = 0.698), a_0 (RMSE = 0.512, R^2 = 0.883), b_0 (RMSE = 0.404, R^2 = 0.458), v_0 (RMSE = 0.289, R^2 = 0.546).

3.2. AbraLlama-Solute

The AbraLlama-Solute model predicts Abraham solvent descriptors (E, S, A, B, V) with high accuracy, achieving R^2-values over 0.9 for all descriptors except A, which has a cross-validated R^2-value of 0.85 (see Table 1). To investigate outliers, we compare the predicted log P values using the 1-octanol solvent parameters (c = 0.088, e = 0.562, s = −1.054, a = 0.034, b = −3.460, v = 3.814) calculated from the measured and predicted solute descriptors. The high density of points along the 45-degree line (see Figure 2) demonstrates the effectiveness of the solute descriptor prediction models and helps identify outliers.

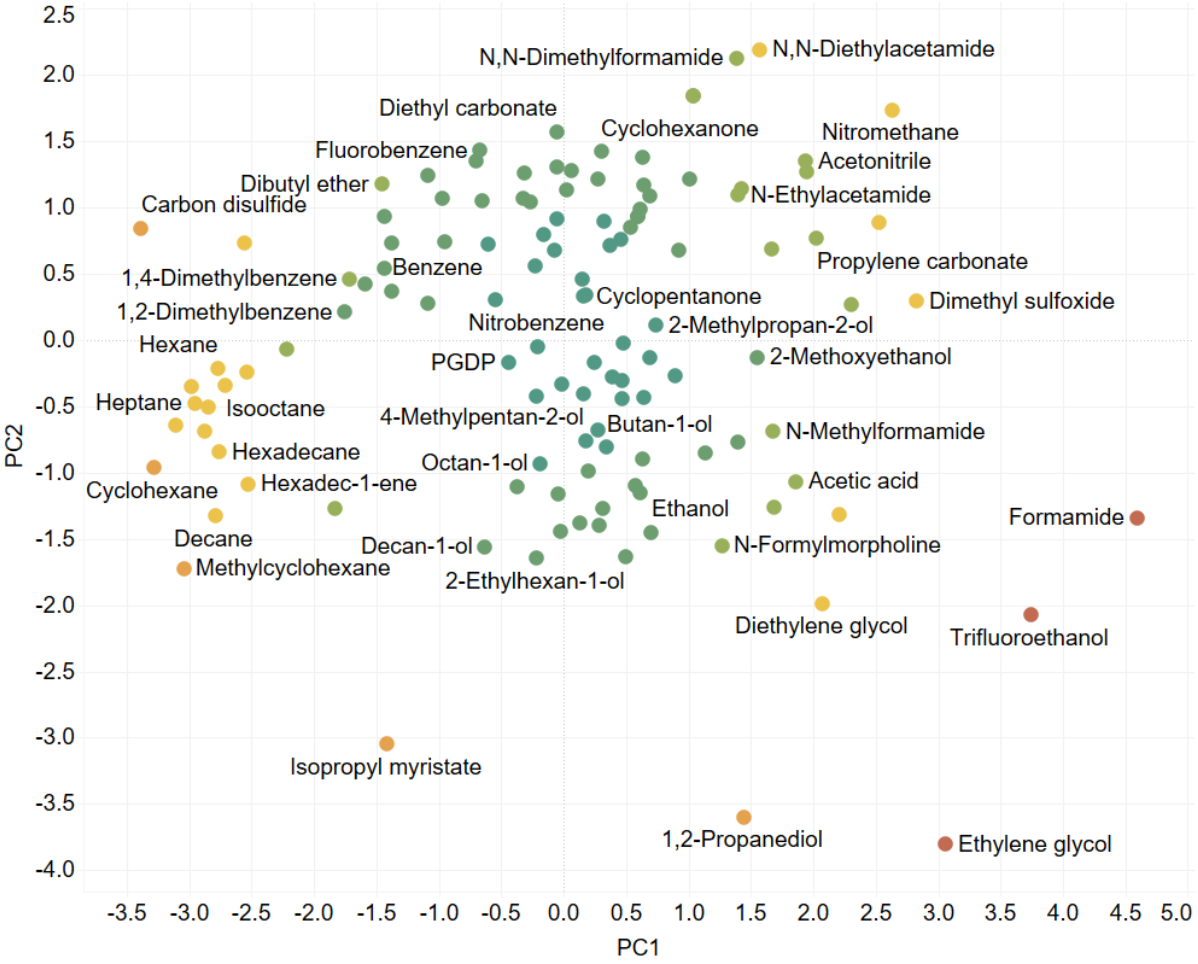


Figure 1. Solvent chemical space (PCA) colored by prediction error: green indicates small errors (Euclidean distance < 1), yellow-orange indicates medium errors (1 ≤ Euclidean distance ≤ 1.45), and red indicates large errors (Euclidean distance > 1.45).

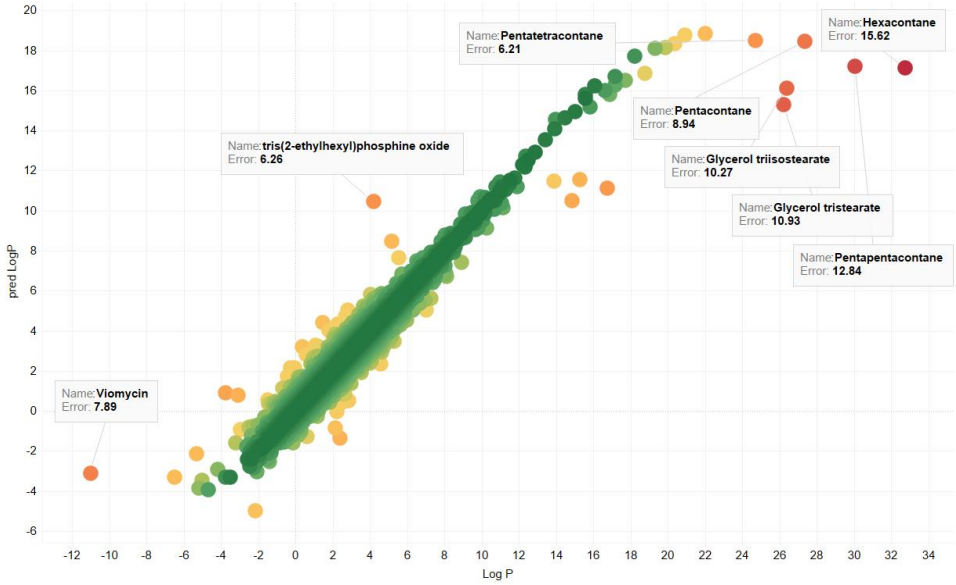


Figure 2. Calculated log P values (N = 6852) using the solvent parameters for 1-octanol, comparing measured vs. predicted Abraham solute descriptors. The data points are colored by absolute error.

4. Conclusions

This study successfully demonstrates the application of fine-tuned large language models (LLMs) in predicting molecular properties, specifically the prediction of Abraham model solute descriptors (E, S, A, B, V) and modified Abraham solvent parameters (e_0, s_0, a_0, b_0, v_0). By leveraging the capabilities of ChemLLaMA [4], we developed the AbraLlama-Solvent and AbraLlama-Solute models, which predict these critical parameters with accuracy consistent with previous studies [5,13].

The AbraLlama-Solvent model predicts modified solvent parameters across a wide range of solvents, facilitating the comparison of solvents with similar solvation properties. The PCA highlighted that prediction errors correlate with the distance from the center of the chemical space, indicating a need for further refinement in regions containing complex solvent structures and for the considered use of Hugging Face apps [12]. Similarly, the AbraLlama-Solute model achieves high accuracy in predicting Abraham solute descriptors, demonstrating its robustness and reliability.

Future work will focus on expanding the solvent dataset to include a broader range of solvents, enhancing the model's accuracy in predicting solvent parameters, and developing a user-friendly solvent selection tool.

Author Contributions: Conceptualization, A.S.I.D.L.; methodology, A.S.I.D.L. and Y.L.; Hugging Face apps, Y.L.; data curation, A.S.I.D.L.; writing, A.S.I.D.L. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Solute data are available for download online [13]. Solvent data are available under a CC0 license from Figshare [14,16].

Acknowledgments: We would like to extend our sincere gratitude to William E. Acree, Jr. for generously providing the solvent data used in this analysis and for his many years of invaluable collaboration and support. His contributions have been instrumental in advancing our research and in fostering an open and collaborative scientific community.

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

References

1. Deng, J.; Yang, Z.; Wang, H.; Ojima, I.; Samaras, D.; Wang, F. A systematic study of key elements underlying molecular property prediction. *Nat. Commun.* **2023**, *14*, 6395. [CrossRef] [PubMed]
2. Lang, A.S.I.D.; Chong, W.K.; Wörner, J.H. Fine-Tuning ChemBERTa-2 for Aqueous Solubility Prediction. *Ann. Chem. Sci. Res.* **2023**, *4*, 1–3. [CrossRef]
3. Luong, K.-D.; Singh, A. Application of Transformers in Cheminformatics. *J. Chem. Inf. Model.* **2024**, *64*, 4392–4409. [CrossRef] [PubMed]
4. Lee, Y.; Lang, A.S.I.D.; Cai, D.; Wheat, S.R. The Role of Model Architecture and Scale in Predicting Molecular Properties: Insights from Fine-Tuning RoBERTa, BART, and LLaMA. *arXiv* **2024**. [CrossRef]
5. Bradley, J.-C.; Abraham, M.H.; Acree, W.E.; Lang, A.S. Predicting Abraham model solvent coefficients. *Chem. Cent. J.* **2015**, *9*, 12. [CrossRef] [PubMed]
6. Chung, Y.; Vermeire, F.H.; Wu, H.; Walker, P.J.; Abraham, M.H.; Green, W.H. Group Contribution and Machine Learning Approaches to Predict Abraham Solute Parameters, Solvation Free Energy, and Solvation Enthalpy. *J. Chem. Inf. Model.* **2022**, *62*, 433–446. [CrossRef] [PubMed]
7. Abraham, M.H.; Zissimos, A.M.; Acree, W.E. Partition of solutes into wet and dry ethers; an LFER analysis. *New J. Chem.* **2003**, *27*, 1041–1044. [CrossRef]
8. Abraham, M.H.; Acree, W.E. Comparison of solubility of gases and vapours in wet and dry alcohols, especially octan-1-ol. *J. Phys. Org. Chem.* **2008**, *21*, 823–832. [CrossRef]
9. Abraham, M.H.; Smith, R.E.; Luchtefeld, R.; Boorem, A.J.; Luo, R.; Acree, W.E. Prediction of solubility of drugs and other compounds in organic solvents. *J. Pharm. Sci.* **2010**, *99*, 1500–1515. [CrossRef] [PubMed]
10. Jouyban, A.; Acree, W.E., Jr. Michael H. Abraham and his developed parameters: Various applications in medicine, chemistry and biology. *Pharm. Sci.* **2022**, *28*, 170–173. [CrossRef]
11. Lee, J.L.; Chong, G.H.; Ota, M.; Guo, H.; Smith, R.L. Solvent Replacement Strategies for Processing Pharmaceuticals and Bio-Related Compounds—A Review. *Liquids* **2024**, *4*, 352–381. [CrossRef]

12. Lang, A.S.I.D.; Lee, Y. AbraLlama Hugging Face App: Predicting Abraham Model Solute Descriptors and Modified Solvent Parameters Using Llama. Hugging Face. 2024. Available online: https://huggingface.co/spaces/ttmn/AbraLlama (accessed on 24 May 2024).
13. Ulrich, N.; Endo, S.; Brown, T.N.; Watanabe, N.; Bronner, G.; Abraham, M.H.; Goss, K.-U. *UFZ-LSER Database v 3.2.1*; Helmholtz Centre for Environmental Research-UFZ: Leipzig, Germany, 2017; Available online: http://www.ufz.de/lserd (accessed on 24 May 2024).
14. Acree, W.E., Jr.; Land, A.S.I.D.; Lee, Y. Dataset: Abraham model Log P and Log K equation coefficients. *Figshare* **2024**. [CrossRef]
15. Sinha, S.; Yang, C.; Wu, E.; Acree, W.E., Jr. Abraham Solvation Parameter Model: Examination of Possible Intramolecular Hydrogen-Bonding Using Calculated Solute Descriptors. *Liquids* **2022**, *2*, 131–146. [CrossRef]
16. Lang, A.S.I.D.; Lee, Y. Dataset: AbraLlama: Predicting Abraham Model Solute Descriptors and Modified Solvent Parameters Using Llama. *Figshare* **2024**. [CrossRef]
17. Lee, Y.; Lang, A.S.I.D.; Cai, D.; Wheat, S.R. Transformers and Chemistry. Available online: https://github.com/BrightBlueCheese/transformers_and_chemistry (accessed on 24 May 2024).
18. Falcon, W. The PyTorch Lightning Team. PyTorch Lightning (Version 1.9.5). 2024. Available online: https://github.com/Lightning-AI/pytorch-lightning/ (accessed on 24 May 2024).
19. The PyTorch Lightning Bolts Team. PyTorch Lightning Bolts (Version 0.7.0). 2024. Available online: https://github.com/Lightning-Universe/lightning-bolts (accessed on 24 May 2024).
20. Lee, Y.; Lang, A.S.I.D. AbraLLaMA Source Code. Available online: https://github.com/BrightBlueCheese/AbraLLaMA (accessed on 24 May 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.