SoDaDE: Solvent Data-Driven Embeddings with Small Transformer Models

Gabriel Kitso Gibberd
Imperial College London
London, UK

Jose Pablo Folch
SOLVE Chemistry
London, UK

Antonio Del Rio Chanona
Imperial College London
London, UK

Abstract

Computational representations have become crucial in unlocking the recent growth of machine learning algorithms for chemistry. Initially hand-designed, machine learning has shown that meaningful representations can be *learnt* from data. Chemical datasets are limited and so the representations learnt from data are generic, being trained on broad datasets which contain shallow information on many different molecule types. For example, generic fingerprints lack physical context specific to solvents. However, the use of harmful solvents is a leading climate-related issue in the chemical industry, and there is a surge of interest in green solvent replacement. To empower this research, we propose a new solvent representation scheme by developing Solvent Data Driven Embeddings (SoDaDE). SoDaDE uses a small transformer model and solvent property dataset to create a fingerprint for solvents. To showcase their effectiveness, we use SoDaDE to predict yields on a recently published dataset, outperforming previous representations. We demonstrate through this paper that data-driven fingerprints can be made with small datasets and set-up a workflow that can be explored for other applications.

1 Introduction

Chemistry needs better representations for molecular modelling [Yang et al., 2019]. Widely-used representations [Landrum, 2006, Probst et al., 2022] capture general molecular features rather than structures and properties relevant to the modelling task. Increasingly complex data-driven representations [Ahmad et al., 2022, Schwaller et al., 2021a] are helping address these issues, however, they are not guaranteed to generalise, especially to unseen or niche chemistry.

As a clear example, we address the issue of solvent representation highlighted by the newly proposed Catechol Benchmark [Boyne et al., 2025]. Most chemical reactions both in industry and academia occur in solvents. Many solvents are volatile organic compounds (VOCs) which contribute to environmental issues, such as smog formation, and health problems including cancer [Jindamanee et al., 2025]. The exploration for greener and safer solvents is ongoing but work is needed to model solvent impact accurately and identify replacements for harmful VOC solvents.

In this work we create a new data-driven fingerprint for solvents. Dubbed Solvent Data-Driven Embeddings (SoDaDE), we identify a small but suitable solvent property dataset and train a transformer model [Vaswani et al., 2017] via data augmentation. We assess performance on the pre-training task and then showcase the usefulness of the new fingerprint by outperforming previous representations on the Catechol Benchmark [Boyne et al., 2025]. We provide data and code [1].

[1] `https://github.com/g-a-b-r-e-a-l/SoDaDE_Solvent_Data-Driven_Embeddings_from_Language_Models`

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

2 Background and Related Work

2.1 Common Chemical Representations

For human understanding, molecular representations are strings of characters, like colloquial names, IUPAC names and Simplified Molecular Input Line Entry System (SMILES) strings [Weininger, 1988]. For machine understanding, molecular simulations and representations are used. Molecular simulations, such as density functional theory (DFT) calculate molecular behaviour, but complexity scales exponentially with size so simulations struggle with large molecules [Goedecker, 1999].

Molecular representations allow machine learning from chemical datasets by communicating a molecule to a machine. Representations have two main categories, molecular graphs [Duvenaud et al., 2015] and rule-based fingerprints, although descriptors, like DFT values are also used [Ahneman et al., 2018]. A molecular graph is a network, with nodes as atoms and edges as bonds. These show close connectivity well but struggle with global structure and usability [Szymkuć et al., 2016]. Rule-based fingerprints are more accessible and often used for chemical property (ECFPs) [Rogers and Hahn, 2010] and yield prediction (DRFPs) [Probst et al., 2022]. These fingerprints one-hot encode each substructure present into a long vector. Consequently, rule-fingerprints communicate structure well but are unspecific and not information dense.

2.2 Related Work

Based on the success of language foundation models [Devlin et al., 2019, Touvron et al., 2023], similar ideas have emerged to attempt the building of chemical foundation models. From this, data-driven molecular fingerprints (DDfps) have risen to address the issues with common chemical representations [Honda et al., 2019]. DDfps are representations learnt by pre-training a neural network, commonly a transformer model, on related data through self-supervised learning. Then the second last layer, before an output is calculated, is used as a representation. If trained correctly, these vectors will contain information [Pratt, 1992] about the molecules and lead to better results when fine-tuned.

ChemBERTa [Chithrananda et al., 2020, Ahmad et al., 2022], based on BERT [Devlin et al., 2019], is an early molecule DDfp trained on 77 million SMILES strings and showed good performance in property prediction. Reaction Fingerprints (RXNFP) [Schwaller et al., 2021a,b] pre-trained on reaction classification and achieved excellent performance on yield prediction. Recently, T5-Chem [Lu and Zhang, 2022] explored pre-training on multiple tasks simultaneously. However, due to the range of molecular types in large datasets, DDfps tend to perform worse on niche prediction tasks.

2.3 Semantic Understanding of Chemical Data

Different transformer architectures [Vaswani et al., 2017]) are used for fingerprint generation and most learn from text, like ChemBERTa and RXNFP, which rely on tokenising SMILES strings. However, chemical datasets often include numerical data so data-types need to be combined to leverage the entire dataset. We pre-train with a similar scheme to Yin et al. [2020], who project text and tabular data into the same embedding space.

3 Data and Method

Boyne et al. [2025] recently proposed a new machine learning benchmark, named the Catechol Benchmark, for solvent selection. They measured the yields of the rearrangement of allyl substituted catechol under different reaction conditions, combining a variety of solvents, temperatures and reaction times. From their relatively poor results, they concluded that the machine learning community needs better solvent representations.

Out of the representations they explored, the solvent descriptors found in Spange et al. [2021] performed the best. This aligns with chemical understanding; solvent properties are known to affect reaction success, so we focus our transformer model's pre-training on this dataset. A few example rows can be found in Table 1.. The solvent-property relationships gained from pre-training should create a better solvent featurisation to improve upon the Catechol Benchmark results.

Solvent	Type	ET(30)	α	β	π^*	SA	SB	SP	SdP	n	δ
n-pentane	alkane	31	0	0	-0.08	0	0.073	0.593	0	1.358	14.5
3-methylpentane	alkane	–	–	–	–	0	0.05	0.62	0	–	–

Table 1: Example data from the Spange solvent dataset. It contains 191 solvents, their type, and molecular properties. Some values are missing, so we leverage the ability of transformers to mask these values during training. The full set of properties are ET(30): Reichardt's polarity parameter, α: hydrogen donating ability, β: hydrogen accepting ability, π^*: Kalmel-Taft polarisability parameter, SA: solvent acidity, SB: solvent basicity, SP: solvent polarisability, SdP: solvent dipolarity, $N_{mol\ cm^{-3}}$: molecular density, n: solvent refractive index, $f(n)$: a function of n quantifying the non-specific solute-solvent interactions, δ: a correction term for polarisability.

The Spange solvent property set contains up to 12 molecular properties of 191 solvents. While a small dataset, it is appropriate as commonly used solvents are limited. We augment the dataset size by creating 'solvent sequences', which is achieved by randomly shuffling property-value pairs in the sequence a maximum number of 12! combinations per solvent. For training, mask tokens were used to cover-up random solvent properties. We learn a model which attends to preceding, property values and predicts the properties covered by the mask token. A causal mask, common in transformer models, was used to ensure the model focussed on tokens before the token to be predicted. Finally, an attention mask was used to hide missing values from the model. A summary of the data processing and training method is illustrated in Figure 1.

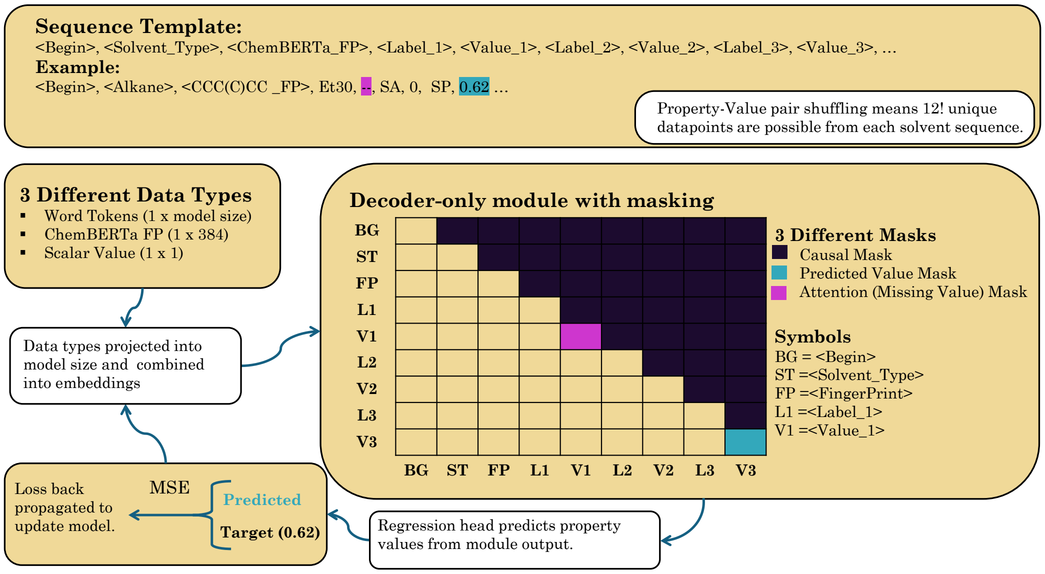
Sequence Template:
<Begin>, <Solvent_Type>, <ChemBERTa_FP>, <Label_1>, <Value_1>, <Label_2>, <Value_2>, <Label_3>, <Value_3>, …
Example:
<Begin>, <Alkane>, <CCC(C)CC _FP>, Et30, --, SA, 0, SP, 0.62 …
Property-Value pair shuffling means 12! unique datapoints are possible from each solvent sequence.
3 Different Data Types
Word Tokens (1 x model size)
ChemBERTa FP (1 x 384)
Scalar Value (1 x 1)
Data types projected into model size and combined into embeddings
Decoder-only module with masking
BG
ST
FP
L1
V1
L2
V2
L3
V3
3 Different Masks
Causal Mask
Predicted Value Mask
Attention (Missing Value) Mask
Symbols
BG = <Begin>
ST =<Solvent_Type>
FP =<FingerPrint>
L1 =<Label_1>
V1 =<Value_1>
Loss back propagated to update model.
MSE
Predicted
Target (0.62)
Regression head predicts property values from module output.

Figure 1: This diagram shows the data augmentation and model training structure. The Spange descriptors (top of the diagram) used to construct solvent sequences which could be shuffled to create more data points. The model combines the items in these sequences into the standardised model size vectors. The decoder then learns from these sequences, predicting any values hidden by the mask token, based on the previous sequence (causal mask) and ignores missing values (attention mask).

Choices within the pre-training task have a significant impact on fine-tuning performance [Liu et al., 2023] so the validation and test set were chosen based on the solvent types in the Catechol dataset. For each of the 9 solvent types in the Catechol dataset, 1 random, complete solvent was taken from the Spange dataset and added to the validation set. This incentivised the transformer model to learn good representations of the relevant solvent types. For the test set, we took 5 additional solvents from the Spange set, one for each of the most common solvent types in the Catechol set to get a proxy of performance on the relevant downstream task.

Model parameters were explored using grid search with the help of a high-performance cluster. Final model parameters for the SoDaDE model were a dimension size of 64, 16 attention heads, 5 layers and a hidden dimension of 4 × 64. A learning rate scheduler was used which started at a learning rate of 0.001 and was reduced by a factor of 0.5 every 5 epochs of no improvement until it reached a learning rate of 1e-8. It used a mask token rate of 0.3 and achieved a final, normalised MSE of 0.107 on the validation set. For fine-tuning, on the Catechol benchmark task, the SoDaDE decoder output of the last token was used as the solvent representation. This was passed through one small multi-layer perceptron (MLP) and then combined with the temperature and residence time in a second MLP to predict the percentage starting material, product 2 and product 3. We test two versions of our model, one where we fix the SoDaDE weights (i.e. without fine-tuning), and another when we allow for changing of the weights when training on the data (i.e. with fine-tuning).

4 Results

4.1 Solvent Property Prediction

While not a guarantee of success, performance on a relevant pre-training task is a good indicator that the model is learning effective representations [Liu et al., 2023]. In Table 2 performance of SoDaDE is compared with a Gaussian process (GP) model, Random Forest (RF) model, and the average of each property in the training set (AVG). For the competing models, we featurize solvents using molecular fingerprints calculated using RDKit [Landrum, 2006, Rogers and Hahn, 2010].

Table 2: This table compares the performance of different methods on predicting properties of the 5 solvents in the test set using MSE.

Metric	Averaged Values	RF Predictions	GP Predictions	SoDaDE Predictions
ET30	40.06	4.00	1.33	**1.22**
α	0.115	**0.0004**	0.0050	0.0010
β	0.0750	0.0237	0.0143	**0.0084**
$\pi^{\star}$	0.112	0.0515	**0.0057**	0.0170
SA	0.0201	0.0001	0.0012	**0.00009**
SB	0.0617	0.0280	0.0144	**0.0020**
SP	0.0083	0.0033	**0.0008**	0.0012
SdP	0.0848	0.0294	**0.0128**	0.0160
$N_{mol\ cm^{-3}}$	4.00e-6	2.00e-6	**1.00e-6**	**1.00e-6**
n	0.0050	0.0013	0.0006	**0.0003**
fn	0.0013	0.0003	0.0002	**0.00007**
δ	10.2	2.46	2.64	**0.771**
Average MSE	4.23	0.550	0.335	**0.170**

4.2 Performance on the Catechol Benchmark

The Catechol Benchmark is split into two different tasks, one of predicting reaction outcomes with a single solvent, and a second "full dataset" task. The full data involves predicting reaction outcomes under multiple solvent mixtures. To account for this, weighted average of the two solvent fingerprints were taken and fed to the neural network. The model was adapted to a fork of the Catechol GitHub to ensure consistent testing. These results are displayed in Table 3.

We compare against the ACS solvent selection guide's PCA representation [Diorazio et al., 2016], the differential reaction fingerprints [Probst et al., 2022], RDKit's molecular fragments and fingerprints concatenated (fragprints) [Landrum, 2006, Griffiths et al., 2022], Spange et al. [2021]'s featurization with GP imputation, reaction fingerprints [Schwaller et al., 2021a,b], and ChemBERTa fingerprints [Ahmad et al., 2022]. These are all the methods compared against in the original benchmarking paper [Boyne et al., 2025]. Interestingly, we achieve stronger performance in the full data as opposed to the single solvent data. This suggests we have learnt a good representation for solvent mixtures, which original methods struggled to do well.

Table 3: Comparison of the performance of different methods on the Catechol benchmark. The competing method performances are taken from the original paper. Their repository was forked and SoDaDE was added to ensure consistency in the task. The values are the MSE between the predicted yield values and actual values of the starting material, and two products as percentages.

Model	Featurisation	MSE Full Data (↓)	MSE Single Solvent (↓)
MLP	ACS	0.0140	0.0110
	DRFPs	0.0130	0.0150
	Fragprints	0.0110	0.0100
	Spange (GP imputation)	0.0100	0.0100
LLM	RXNFP	0.1050	0.0550
	ChemBERTa	0.1530	0.0740
	SoDaDE (ours) without tuning	0.0029	**0.0044**
	SoDaDE (ours) with tuning	**0.0026**	**0.0044**

5 Conclusion

The SoDaDE model shows significant improvements over tested methods, demonstrating that we have created an effective solvent fingerprint. The similarity between the non-finetuned and fine-tuned models shows this fingerprint is not dependant on fine-tuning and captures general solvent features. Through SoDaDE, we demonstrate that effective fingerprints can be created from small datasets and invite others to replicate our method within their own domains.

References

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(8): 3370–3388, 2019.

Greg Landrum. RDKit: Open-source cheminformatics., 2006. URL `https://www.rdkit.org`.

Daniel Probst, Philippe Schwaller, and Jean-Louis Reymond. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digital discovery*, 1(2):91–97, 2022.

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa-2: Towards chemical foundation models. *2021 ELLIS Machine Learning for Molecule Discovery Workshop*, 2022.

Philippe Schwaller, Daniel Probst, Alain C Vaucher, Vishnu H Nair, David Kreutter, Teodoro Laino, and Jean-Louis Reymond. Mapping the space of chemical reactions using attention-based neural networks. *Nature machine intelligence*, 3(2):144–152, 2021a.

Toby Boyne, Juan S Campos, Becky D Langdon, Jixiang Qing, Yilin Xie, Shiqiang Zhang, Calvin Tsay, Ruth Misener, Daniel W Davies, Kim E Jelfs, et al. The catechol benchmark: Time-series solvent selection data for few-shot machine learning. *Advances in Neural Information Processing Systems*, 38, 2025.

Kanisorn Jindamanee, Jutarat Keawboonchu, Nattaporn Pinthong, Aronrag Meeyai, Puchong Inchai, and Sarawut Thepanondh. Environmental impacts and emission profiles of volatile organic compounds from petroleum refineries. *Scientific Reports*, 15(1):15509, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Stefan Goedecker. Linear scaling electronic structure methods. *Reviews of Modern Physics*, 71(4): 1085, 1999.

David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28, 2015.

Derek T Ahneman, Jesús G Estrada, Shishi Lin, Spencer D Dreher, and Abigail G Doyle. Predicting reaction performance in c–n cross-coupling using machine learning. *Science*, 360(6385):186–190, 2018.

Sara Szymkuć, Ewa P Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald, Michał Startek, Michał Bajczyk, and Bartosz A Grzybowski. Computer-assisted synthetic planning: the end of the beginning. *Angewandte Chemie International Edition*, 55(20):5904–5937, 2016.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Shion Honda, Shoi Shi, and Hiroki R Ueda. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738*, 2019.

Lorien Y Pratt. Discriminability-based transfer between neural networks. *Advances in neural information processing systems*, 5, 1992.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. *Machine Learning for Molecules Workshop at NeurIPS 2020*, 2020.

Philippe Schwaller, Alain C Vaucher, Teodoro Laino, and Jean-Louis Reymond. Prediction of chemical reaction yields using deep learning. *Machine learning: science and technology*, 2(1): 015016, 2021b.

Jieyu Lu and Yingkai Zhang. Unified deep learning model for multitask reaction predictions with explanation. *Journal of chemical information and modeling*, 62(6):1376–1387, 2022.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. TaBERT: Pretraining for joint understanding of textual and tabular data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Stefan Spange, Nadine Weiß, Caroline H Schmidt, and Katja Schreiter. Reappraisal of empirical solvent polarity scales for organic solvents. *Chemistry-Methods*, 1(1):42–60, 2021.

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. Same pre-training loss, better downstream: Implicit bias matters for language models. In *International Conference on Machine Learning*, pages 22188–22214. PMLR, 2023.

Louis J Diorazio, David RJ Hose, and Neil K Adlington. Toward a more holistic framework for solvent selection. *Organic Process Research & Development*, 20(4):760–773, 2016.

Ryan-Rhys Griffiths, Jake L Greenfield, Aditya R Thawani, Arian R Jamasb, Henry B Moss, Anthony Bourached, Penelope Jones, William McCorkindale, Alexander A Aldrick, Matthew J Fuchter, et al. Data-driven discovery of molecular photoswitches with multioutput Gaussian processes. *Chemical Science*, 13(45):13541–13551, 2022.

A Investigation of SoDaDE parameters

Deep neural networks are models that lack significant interpretability. They capture complex relationships in a way that is difficult to extract, visualise, or understand. This lack of interpretability is a loss for chemistry, where the relationships found could contribute significantly to chemical understanding.

Figure 2 plots the solvent embeddings from the SoDaDE and NN models over 200 batches of training. These embeddings have been reduced from 64-bit vectors to 2 bits using principal component analysis (PCA). The colour map values use the conversion efficiency, Equation 1, of the catechol reaction in each solvent. To calculate conversion efficiency, the maximum quantities of Products 2 and 3 and the minimum quantity of SM were used. Colour mapping this way achieves better distinction between solvents.

$$Value = \frac{(Product2 + Product3)}{(Product2 + Product3 + SM)} \tag{1}$$

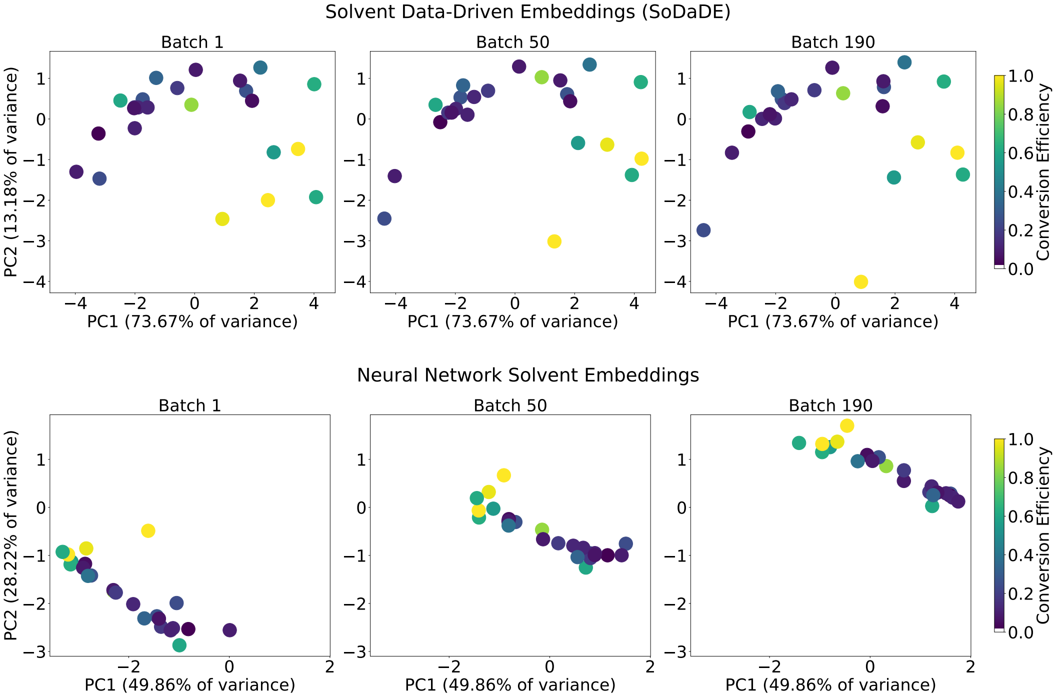

Figure 2: This shows the transformation of the solvent embeddings during training, plotted in PCA-reduced space from 64 dimensions down to 2. The neural network (NN) embeddings organise as training continues, moving from the lower left corner to the upper right. The SoDaDE embeddings are disordered and do not organise noticeably. The low learning rate of the SoDaDE model likely contributes to this, but the performance achieved suggests the model creates effective solvent fingerprints.