Systems & Control Transactions

Original Research Article
Peer Reviewed Conference Proceeding
ESCAPE 36 - European Symposium on Computer Aided Process Engineering
Sheffield, UK. 21-24 June 2026

PSE PRESS

Optimizing the Solubility of Organic Molecules in Mixed Solvents Using Bayesian Optimization and Multicomponent Directed-Message Passing Neural Networks

Simona Buzzi[a], Ulderico Di Caprio[a], Dominik Bongartz[a], and Florence Vermeire[a]*

[a] KU Leuven, Department of Chemical Engineering, Celestijnenlaan 200F, Leuven 3001, Belgium.
* Corresponding Author: florence.vermeire@kuleuven.be

ABSTRACT

Accurate prediction of solubility limits of organic compounds in mixed solvents is critical for the design and optimization of chemical and pharmaceutical processes. Recent advances in machine learning have enabled fast and reliable prediction of physicochemical properties of molecules, including solubility. In this work, we present a Bayesian Optimization framework to identify optimal solvent combinations, compositions, and temperatures that maximize the solubility of active pharmaceutical ingredients. The optimization strategy leverages a multicomponent directed-edge message passing neural network trained on solvent mixtures to predict solubility in ternary systems consisting of a solute and two solvents. To enable efficient Bayesian optimization, we represented the solvents in a continuous space and compare three different strategies: integer enumeration, numerical descriptors, and deep embeddings. The proposed approach was tested on a dataset comprising 14 299 points in solvent mixtures. The results indicate that integer enumeration and embedding-based representations achieve the best performance in identifying solvent mixtures that maximize solubility for each compound at a fixed number of model evaluations. While demonstrated on ternary systems, the framework is readily extensible to mixtures with a larger number of components. With this work we aim to identify which chemical representation is most suitable for similar optimization problems, providing a solid foundation for further and broader exploration, for example in the optimization of crystallization processes.

Keywords: Bayesian Optimization, Deep Learning, Mixed Solvents, Solubility

INTRODUCTION

Predicting solubility limits of organic compounds in mixed solvents is crucial in various fields, including environmental science, chemical engineering and pharmaceuticals. In the pharmaceutical industry, for instance, the synthesis of Active Pharmaceutical Ingredients (APIs) is gradually moving from traditional batch processes to processes in continuous flow, due to their enhanced safety and sustainability [1, 2]. However, one of the main hurdles of continuous flow reactors is the precipitation of solids within the reactor, leading to equipment clogging and severe safety risks. Therefore, controlling the solubility of reaction intermediate species and products is crucial for the design and optimization of such processes. Moreover, information on solubility limits is equally important in the stages following the synthesis, such as crystallization, as well as in telescoped reaction sequences.

Traditionally, solubility is calculated using thermodynamic cycles of fusion or sublimation. When considering a fusion cycle, the solubility can be calculated using the definition of ideal solubility which is expressed in terms of the fusion properties of the solute, namely the melting temperature and the enthalpy of fusion [3]. The non-ideality is accounted through activity coefficients (γ). Several approaches are available to estimate γ, including the Wilson equation of state [4] and the Non-Random Two-

Liquid (NRTL) model [5]. However, these models rely on interaction parameters that must be regressed from vapor–liquid equilibrium data, which are often unavailable for many systems of practical interest. In the absence of such parameters, group contribution methods, including UNIFAC [6] can be employed. Methods from quantum chemistry, including both implicit and explicit solvent models, can also be used to compute activity coefficients. Nevertheless, these approaches are computationally demanding, as they require extensive geometry optimization. Moreover, such methods often do not account for the presence of a solid phase, limiting their applicability for accurate solubility predictions. More recently, purely data-driven techniques, including quantitative structure-property relationship (QSPR) models [7] and machine learning (ML) [8] approaches have been promising in predicting molecular properties. In particular, ML has emerged as a valuable alternative to traditional modeling approaches in chemical engineering applications, enabling fast and accurate predictions of complex physicochemical properties.

Algorithms for solvent optimization are closely tied to the accuracy of the predictive models they rely on. Although several studies exist, most of them employ group-contribution approaches as predictive models [9]. In our recent work, we developed robust deep learning models for solubility prediction in mixtures of organic solvents [10]; thanks to these accurate and predictive models, it is now possible to develop optimization strategies to identify solvent mixtures and experimental conditions that optimize solubility. Such tools enable a more efficient exploration of solvent space, reduce the experimental costs and improve the process efficiency. The main limitation of using these ML models is the computational cost associated with their evaluation; direct evaluation of all possible solvent combinations with predictive deep learning models is often computationally prohibitive due to the combinatorial explosion of possible mixtures, especially when considering multiple solvents beyond binary systems.

Bayesian Optimization (BO) [11] is an effective approach for optimizing costly objective functions when their analytical form is unknown or difficult to evaluate, as is the case with graph neural networks. In BO, a surrogate model, often a Gaussian process (GP), is constructed to approximate the expensive objective function. This surrogate is then used to guide the selection of new evaluation points through an acquisition function, which balances exploration and exploitation. By leveraging the acquisition function, BO can efficiently identify maxima or minima with a limited number of function evaluations. A key strength of this approach is its ability to handle both continuous and discrete variables.

In this work, we present an optimization tool based on BO to identify optimal solvent combinations, compositions, and temperatures that maximizes the solubility of APIs. The method leverages a multicomponent directed-edge message passing neural network (D-MPNN) [12] to predict the solubility in ternary systems, i.e., an API dissolved in two solvents [10]. BO is employed to efficiently explore the solvent design space by constructing a GP surrogate model of the mixture D-MPNN, using a radial basis function kernel defined over continuous input variables. Since the molecular identifiers for the solvents cannot be directly incorporated into the Gaussian process kernel, three representation strategies were adopted: integer labels, molecular descriptors, and deep embeddings. This study primarily investigates the selection of chemical representations for the optimization algorithm, aiming to develop a fully automated digital framework that allows users, given a target molecule and a set of solvents, to determine the optimal conditions for pharmaceutical substance design, from chemical synthesis to purification.

BACKGROUND

Directed-Message Passing Neural Networks

D-MPNNs are a variant of message passing neural networks (MPNNs) [13] in which, in contrast to standard MPNNs, each bond is represented as two directed edges, allowing messages to flow explicitly from a source atom to a target atom. For each directed edge, a message is computed in a message passing function M_t (eq. 1) based on the features of the source atom, neighboring atoms, and the connecting bond. These messages update the hidden states of the edges through learned transformations and non-linear activations (e.g., ReLU) (eq.2). The initial hidden state for each edge is created by concatenating the features of the source atom and the directed bond. This combined vector is transformed using a learned weight matrix and an activation function (eq.3). The final hidden state for a certain atom, h_v, is computed by using the final message, m_v, obtained by summing all the messages from the previous steps (eq.4). The final message is then used together with the initial atom features, x_v, to give a representation of the atom. Then all the atom hidden states are aggregated, e.g. summed, and a feature vector for the molecule is obtained (eq.5).

$$m_{vw}^{t+1} = \sum_{k \in N\{(v)\setminus w\}} M_t\,(x_v, x_k, e_{vw}) \quad (1)$$

$$h_{vw}^{t+1} = U_t(h_{vw}^{t}, m_{vw}^{t+1}) = \tau(h_{vw}^{0}, + W_m\, m_{vw}^{t+1}) \quad (2)$$

$$h_{vw}^{0} = \tau(W_i\, cat\,(x_v, e_{vw})) \quad (3)$$

$$m_v = \sum_{k \in N(v)} h_{kv}^{T} \quad (4)$$

$$h = \sum_{v \in G} h_v \quad (5)$$

Finally, the molecular vector is passed through a neural

network to predict the properties of interest.
The D-MPNN architecture is implemented in Chemprop [12], one of the most widely used software tools in chemistry.

Bayesian Optimization

Bayesian Optimization (BO) is a probabilistic framework for the global optimization of functions $f(\mathbf{x})$ that are costly to evaluate or lack analytical expressions. The primary goal of BO is to find a value of $\mathbf{x}^*$, which is the globally optimal value of $\boldsymbol{x}$, that maximizes the objective function $f(\mathbf{x})$ over a predefined domain $\mathcal{X} \subset \mathbb{R}^d$

$$\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) \tag{6}$$

In some cases, evaluating $f(\mathbf{x})$ requires substantial computational resources or experimental effort. Moreover, BO is particularly useful when the objective function is treated as a black box, meaning that only function evaluations are available while derivatives are either inaccessible or difficult to compute (e.g., D-MPNN models). Bayesian optimization addresses this challenge by constructing a probabilistic surrogate model that approximates the objective function, often a Gaussian Process, while quantifying uncertainty. A GP is defined by a mean function $\mu(\mathbf{x})$ and a covariance kernel $k(\mathbf{x}, \mathbf{x}')$

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{7}$$

Given a set of observations $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $y_i = f(\mathbf{x}_i) + \epsilon$ and $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ represents observational noise, the GP posterior at a new input $\mathbf{x}_*$ is Gaussian with mean and variance

$$\begin{aligned}\mu_n(\mathbf{x}_*) &= \mathbf{k}_*^\top (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y}, \sigma_n^2(\mathbf{x}_*) \\ &= k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{k}_*\end{aligned} \tag{8}$$

where $\mathbf{K}$ is the covariance matrix of training points and $\mathbf{k}_*$ is the covariance vector between $\mathbf{x}_*$ and the training points. This probabilistic prediction allows BO to identify regions of high uncertainty and guide exploration efficiently. The next point to evaluate is chosen by maximizing an acquisition function $\alpha(\mathbf{x}; \mathcal{D}_n)$, which balances exploration and exploitation. Common acquisition functions include Expected Improvement (EI), Probability of Improvement (PI), and Upper Confidence Bound (UCB), each combining the GP mean and variance in different ways to guide the search. The BO procedure is iterative: the surrogate is updated with new observations, the acquisition function is optimized to propose the next evaluation point, and the objective function is sampled. Finally, this loop continues until convergence, or a maximum number of evaluations is reached. By efficiently leveraging both the predicted mean and uncertainty, Bayesian Optimization reduces the number of expensive function evaluations, making it particularly suited for applications where computational or experimental costs are significant.

METHODOLOGY

Solvent Feature Encoding

In this section we describe the three molecular representations employed for the BO kernel: integer enumeration, numerical descriptors and deep embeddings. For the integer enumeration, all unique solvents and solutes were assigned discrete numerical identifiers. Specifically, the solvent SMILES were extracted, mapped to integer values starting from one and stored in a dictionary, where the keys correspond to the SMILES strings and the values to the assigned integers. The second encoding method employs molecular descriptors. Numerical descriptors were computed using RDKit (version 2024.09.6). Each solvent SMILES was converted into an RDKit mol object, from which a set of physicochemical and structural descriptors was calculated. These included molecular weight, LogP, topological polar surface area, numbers of hydrogen bond donors and acceptors, number of rotatable bonds, ring count and fraction of sp^3-hybridized carbons. The descriptors were manually selected to capture relevant chemical properties for solubility. Finally, solvent embeddings were generated using the multicomponent directed edge graph neural network trained on single solvent systems and mixtures. To ensure consistency, embeddings were extracted using the same pre-trained model employed as the objective function in the optimization framework (see "Solubility predictor"). The latent representation of the solvents was extracted from the solvent-specific message-passing block of the D-MPNN. For dimensionality reduction and visualization, principal component analysis (PCA) from scikit learn (version 1.7.1) was applied, retaining the first 5, 10, 20, and 30 principal components for comparison. The reduced representations were subsequently used both for visualization and exported in JSON format for downstream analysis.

Solubility Predictor

We employ a D-MPNN extended to multicomponent solvent systems implemented in Chemprop (version 2.1.2) as objective function. Molecular graphs of the solute and solvents are encoded using a shared bond-centered D-MPNN in which messages are defined on directed bonds and propagated for three message-passing steps while excluding reverse-bond information. Initial bond messages are constructed from default atom and bond features (86-dimensional) and projected into a latent space of dimension 300. Message updates are performed using learned linear transformations with ReLU activations, resulting in 300-dimensional directed bond embeddings. After 3 message passing steps, bond representations are aggregated to atom-level embeddings

and pooled using mean aggregation to obtain fixed-size 300-dimensional molecular embeddings for each component. Solvent embeddings are then combined using a mixture aggregation module that incorporates solvent mole fractions in a permutation-invariant manner, yielding a 300-dimensional mixture embedding. The solute and mixture embeddings are concatenated (602 dimensions in total, including global features) and passed to a feed-forward neural network with a single hidden layer of size 300 to predict solubility.

An ensemble of four models was trained on logarithmic solubility data (logS [mol/L]) in binary and ternary systems from the datasets BiggerSolDB [14] and MixtureSolDB [15] respectively. The combined dataset comprises 227 998 datapoints, 9 690 unique solutes, and 132 unique solvents in the range of 243.15-592.9K. Training was performed for 100 epochs with a batch size of 256 for training and 2048 for validation and testing. The mean squared error (MSE) was used as the loss function. The dataset was randomly split into training, validation, and test sets with ratios of 0.89, 0.10, and 0.01, respectively. The Chemprop default optimizer (Adam) and learning rate were used, and the predictive performance of the model was evaluated on a held-out test set. Figure 1 shows a schematic representation of the multicomponent D-MPNN architecture.

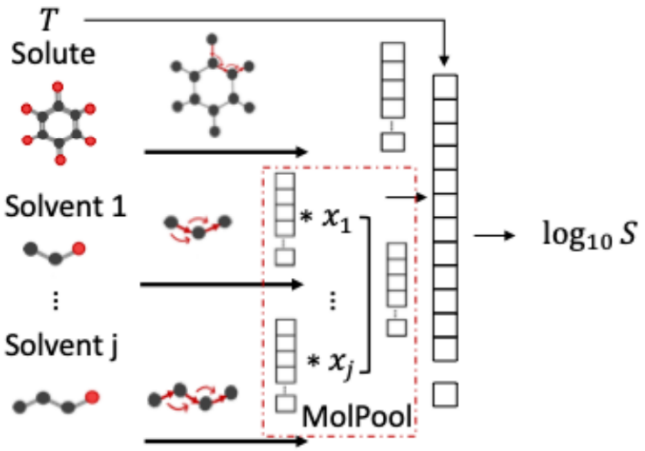

Figure 1. schematic representation of the D-MPNN architecture employed in this work as the objective function within the optimization framework.

Optimization algorithm

As the optimization algorithm, gp_minimize from the scikit-optimize library (version 0.10.2) was employed. Regarding the search space, different bounds were adopted for the categorical variables depending on the solvent representation (i.e., integers, descriptors, or deep embeddings). In the case of integer enumeration, each solvent variable was represented as a discrete integer corresponding to a unique solvent in the dataset. The search space was therefore discrete, ranging from 1 to the total number of unique solvents.

For descriptors and deep embeddings, the search space was instead constructed directly in the continuous representation space of the solvents. Optimized continuous solutions in the descriptor or embedding space were subsequently mapped back to real solvents using a nearest-neighbor projection, selecting the solvent whose representation is closest in Euclidean distance to the optimized point, without modifying the kernel.

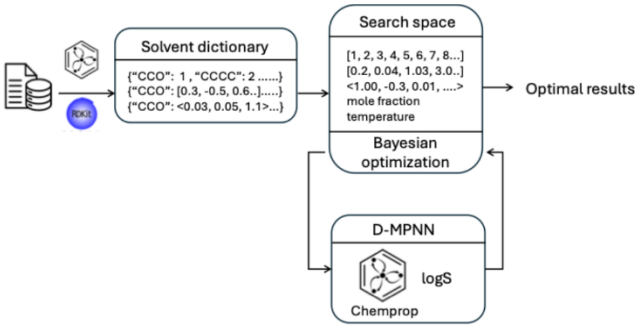

Figure 2. Overview of the optimization workflow: Solvents are extracted from the dataset and encoded as random integers, deep embeddings, or numerical descriptors, which are stored in dictionaries. Bayesian optimization iteratively searches for the best combinations, which are evaluated by the predictor. This loop continues until convergence or a fixed number of iterations is reached.

For all three optimization approaches, the mixture composition was encoded as a continuous variable. The temperature was also modeled as a continuous variable, defined over a real-valued domain bounded by the minimum and maximum temperatures observed in the dataset. The boiling temperature of each pure solvent was estimated using the Joback group contribution method [16]. These boiling temperatures were encoded as numerical values and used together with the solvent representations (integer encoding, molecular descriptors, or embeddings). These values were implemented as a hard-coded constraint in the optimization algorithm, ensuring that solvent mixtures were not proposed at temperatures exceeding the boiling point of the individual pure solvents. This constraint represents an approximation, as the pure component boiling point does not necessary represent the one of the mixtures. Future work could improve this aspect by incorporating predictive models for bubble and dew points. Additionally, to prevent the optimization algorithm from proposing solvent combinations that are implausible from a miscibility standpoint, a constraint based on the water–ethanol partition coefficient was introduced. Future work will focus on extending this framework by incorporating more accurate thermodynamic models for miscibility. Finally, the BO procedure was initialized with 10 random samples and employed

expected improvement as acquisition function. Figure 2 shows an overview of the optimization tool.

RESULTS

The performance of the optimization framework was evaluated on a dataset consisting of 14 299 data points, 66 solutes and 38 solvents in the temperature range 263-367 K. Before reporting the results of the optimization algorithm, we evaluate the performance of the D-MPNN predictive model for the solubility of these solutes. The performance of the model to predict solubility in mixtures of organic solvents for a broader set of solutes and solvents was already extensively evaluated in prior work. In Figure 3, a parity plot shows the performance on the solubility of the 66 tested solutes in the experimentally reported solvents mixtures. Also, for these solutes, a good performance is achieved, indicating that the predicted solubility in new, optimized solvents can be expected to be reliable. Nevertheless, further experimental validation is needed to support these conclusions.

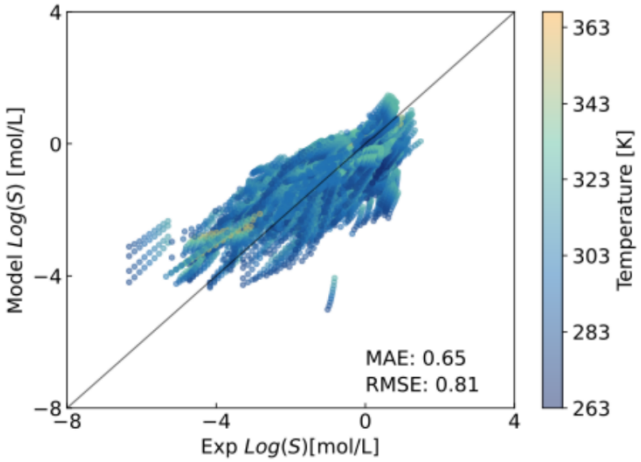

Figure 3. performance of the D-MPNN evaluated on the dataset used in this work, which comprises 66 solutes, 38 solvents, and a total of 14 299 datapoints.

As an example of the performance of the optimization algorithm, we report mixtures proposed by the tool for the three encoding strategies that yield high predicted solubilities. For the enumeration approach, the tool reaches a predicted solubility of 1.4950 $\log_{10}$ mol·L^{-1} for the compound octadecane in a mixture of 4-methylpentan-2-one (0.1 mole fraction) and hexane (0.9) at 333.26 K. This solution is chemically feasible, as both solvents are known to be suitable for octadecane, and the temperature is below their boiling points. The dataset used in this study contains experimentally measured solubilities for only a limited number of solvent mixtures. For octadecane, available experimental data are limited to mixtures of ethanol and cyclohexane. The highest experimentally measured solubility for octadecane was obtained in a mixture of ethanol (0.4) and cyclohexane (0.6) at 297.45 K, with a value of 0.8506 $\log_{10}$ mol·L^{-1}. For the descriptors-based approach, we report the example of propylparaben in a mixture of ethyl acetate (0.2) and dimethylformamide (0.8) at 318.39 K, which was found to have the highest solubility among all solute–solvent combinations. This mixture is feasible because the polar solvents are compatible with the polarity of propylparaben. The optimization algorithm predicted a solubility of 0.8487 $\log_{10}$ mol·L^{-1} for the compound, exceeding the highest experimentally measured solubility in the dataset (0.6249 $\log_{10}$ mol·L^{-1} in pure methanol at 298 K). This demonstrates that the use of solvent mixtures can enhance solubility compared to single-solvent systems. Using the embeddings approach, the maximum predicted solubility (1.6623 $\log_{10}$ mol·L^{-1}) was found for octadecane in a mixture of hexane (0.25) and n-heptane (0.75) at 338.7 K, using an embedding size of 5; since this embedding size yielded the best performance, only results for this setting are reported. The temperature is below the boiling points of both solvents, making the mixture chemically appropriate for octadecane, a paraffin. These results highlight that the embeddings-based approach successfully identified the optimal solvent combination for octadecane, outperforming both other encoding schemes (the descriptor-based model reaches 0.5333 $\log_{10}$ mol·L^{-1}) and the experimental data.

In Figure 4, we present two examples, fumaric acid and 5-nitro-8-hydroxyquinoline, for which we compare the performance of the three representations. In the first case, the enumeration approach achieves higher predicted solubility than both the experimental value and the two encoding schemes. In contrast, for 5-nitro-8-hydroxyquinoline, the chemistry-informed representations (deep embeddings and descriptors) yield better results. For fumaric acid, the enumeration procedure predicts the highest solubility (0.8273 $\log_{10}$ mol·L^{-1}) in a mixture of ethyl acetate (0.1) and dimethylformamide (0.9) at 324.24 K. Descriptors suggest best solubility (-0.0355 $\log_{10}$ mol·L^{-1}) to be in a mixture of diethylene glycol monoethyl ether (0.6) and n-methylpyrrolidone (0.4) at 327.1 K. Finally, for the deep embeddings, the optimal mixture corresponds to diglyme (0.1) and n-methylpyrrolidone (0.9) at 336.97 K (0.0625 $\log_{10}$ mol·L^{-1}). The highest experimentally measured solubility for this molecule, fumaric acid, was obtained in a methanol (0.8) and water (0.2) mixture at 333 K, 0.0898 $\log_{10}$ mol·L^{-1}.

For the other compound, 5-nitro-8-hydroxyquinoline, the best mixture predicted by the enumeration procedure consists of acetone (0.2) and cyclohexanone (0.8) at 300.94 K, yielding −0.2633 $\log_{10}$ mol·L^{-1}. In contrast, the deep embeddings suggest 2-butanone (0.05) and cyclopentanone (0.95) at 324.2 K, with a predicted solubility of 0.1172 $\log_{10}$ mol·L^{-1}. Finally, the descriptor-based model predicts the highest solubility (−0.1440

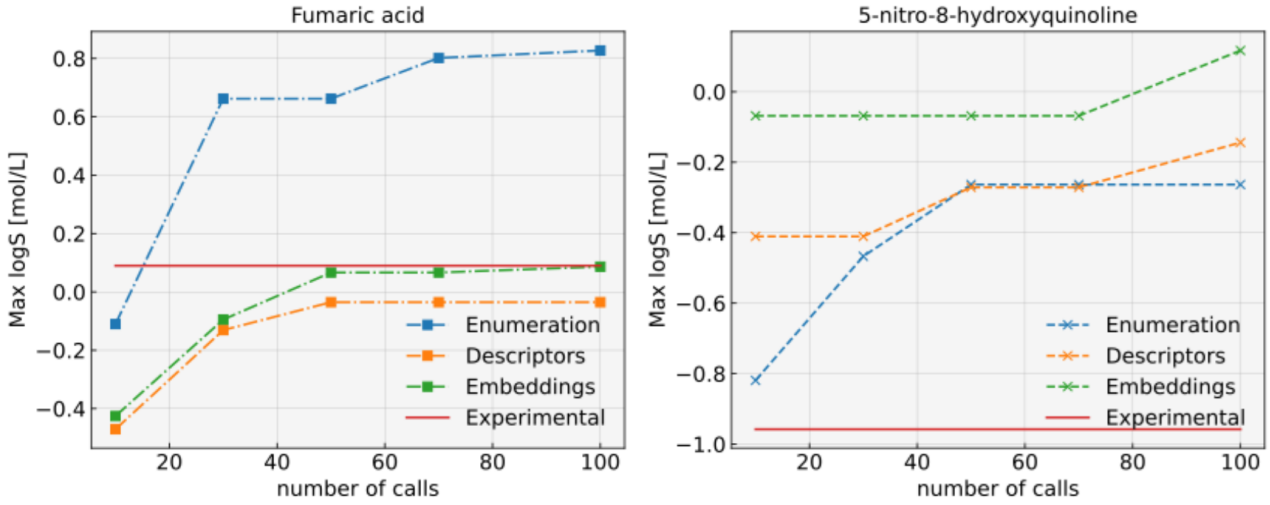

Figure 4. Maximum solubility as a function of the number of optimization calls for the three representations, shown for fumaric acid and 5-nitro-8-hydroxyquinoline. For fumaric acid, integer enumeration and embeddings yield the best performance, whereas for 5-nitro-8-hydroxyquinoline all three representations achieve solubility values higher than the experimental reference. The red continuous line indicates the highest experimentally measured solubility for fumaric acid and 5-nitro-8-hydroxyquinoline.

$\log_{10}$ mol·L^{-1}) in a mixture of pyridine (0.35) and propionic acid (0.65) at 307.5 K. The highest experimental value was -0.9578 $\log_{10}$ mol·L^{-1} found for the mixture ethyl acetate (0.77) and methanol (0.33) at the temperature of 313.15K.

We further analyzed the maximum solubility achieved for each compound in the original curated experimental dataset and compared it with the maxima predicted by the optimization algorithm. Under the enumeration scheme, the algorithm proposed improved solvent mixtures for 36 out of 66 solutes, indicating that for more than half of the compounds it can suggest solvent mixtures that improve solubility as compared to the best experimental measurement. For the embeddings-based approach, 33 solutes exhibited higher predicted solubility than in the original experimental dataset, whereas for the descriptors-based approach, 21 out of 66 solutes showed improved solubility (see table 1). Table 1 summarizes the results for the three enumeration schemes in terms of improvement with respect to the experimental data, mean gain, and mean loss.

These results highlight several important points. First, the enumeration and embedding approaches are the most effective at identifying solvent mixtures that maximize solubility. Second, the proposed method is computationally efficient and improve solubility for a substantial subset of solutes, demonstrating their practical utility for experimental planning. In practice, this approach can significantly reduce the cost associated with testing large numbers of solvent mixtures experimentally. Third, all predicted solutions are chemically reasonable, as they account for solute–solvent, solvent-solvent compatibility and safe temperature ranges.

Table 1: Summary of the three molecular representations in terms of improvement relative to experimental data.

Representation	Better than experiments	Worse than experiments	Mean gain	Mean loss
Enumeration	36	30	0.4881	-0.654
Descriptors	21	45	0.5628	-0.644
Embeddings	33	33	0.5554	-0.599

CONCLUSIONS

In this work, we developed an optimization workflow based on directed message passing neural networks to identify, for each solute, the optimal combination of solvents, mole fractions, and temperatures that maximize solubility. Three different solvent representations were investigated: integer enumeration, physicochemical descriptors, and learned deep embeddings. Our results show that, for a curated dataset with a limited number of solvents and solutes, enumeration and embedding-based representations are the most effective at identifying solvent mixtures that maximize solubility. Both the enumeration strategy and the embeddings-based approach were able to propose improved solvent mixtures for 36 out of 66 solutes and 33 out of 66 solutes whereas the descriptors-based approach led to improved

solubility for a smaller subset of compounds (21 out of 66). To ensure physically meaningful predictions, solvent constraints were imposed to avoid phase separation and to restrict solutions to chemically reasonable temperature ranges. These constraints are necessary because current predictive models do not account for phase stability and lack inherent temperature constraints. However, they should be viewed as approximations that may ultimately be replaced by models that directly enforce thermodynamic constraints within the architecture. Ultimately, the proposed approach lays the groundwork for a fully automated optimization pipeline capable of identifying optimal experimental conditions for organic molecules, spanning the entire workflow from synthesis to crystallization. Future directions include modifying the kernel to directly incorporate physicochemical information, developing improved models for boiling point estimation, enabling the use of optimization algorithms that leverage the structure of the ANN model (e.g., gradient-based methods or mixed-integer linear programming), and experimental validation of the proposed improvements.

ACKNOWLEDGEMENTS

S.B acknowledges the Fonds Wetenschappelijk Onderzoek (FWO) for funding (G021924N). U.D.C acknowledges the Fonds Wetenschappelijk Onderzoek (FWO) for funding (12A4D26N). Resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government.

AUTHOR IDENTIFIERS

Author ORCIDs:
Simona Buzzi: 0009-0007-9521-7921
Ulderico Di Caprio: 0000-0001-5194-8721
Dominik Bongartz: 0000-0003-1790-0235
Florence Vermeire: 0000-0002-5607-8152

REFERENCES

1. Plutschack MB, Pieber B, Gilmore K, Seeberger PH. The hitchhiker's guide to flow chemistry. Chem. Rev. 117:11796-11893 (2017). https://doi.org/10.1021/acs.chemrev.7b00183
2. Ince MC, Benyahia B, Vilé G. Sustainability and techno-economic assessment of batch and flow chemistry in seven industrial pharmaceutical processes. ACS Sustainable Chem. Eng. 13:2864-2874 (2025). https://doi.org/10.1021/acssuschemeng.4c09289
3. Nordström FL, Rasmuson ÅC. Prediction of solubility curves and melting properties of organic and pharmaceutical compounds. European Journal of Pharmaceutical Sciences 36:330-344 (2009). https://doi.org/10.1016/j.ejps.2008.10.009
4. Sadeghi M, Rasmuson ÅC. On the estimation of crystallization driving forces. CrystEngComm 21:5164-5173 (2019). https://doi.org/10.1039/c9ce00747d
5. Wilson, G. M. (1964). A new expression for the excess free energy of mixing. J. Am. Chem. Soc., 86, 127.
6. Renon, H., & Prausnitz, J. M. (1969). Estimation of parameters for the NRTL equation for excess Gibbs energies of strongly nonideal liquid mixtures. Industrial & Engineering Chemistry Process Design and Development, 8(3), 413-419.
7. Fredenslund, A., Jones, R. L., & Prausnitz, J. M. (1975). Group-contribution estimation of activity coefficients in nonideal liquid mixtures. AIChE Journal, 21(6), 1086-1099.
8. Katritzky AR, Lobanov VS, Karelson M. QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. Chem. Soc. Rev. 24:279 (1995). https://doi.org/10.1039/cs9952400279
9. Leenhouts RJ, Morgan N, Al Ibrahim E, Green WH, Vermeire FH. Pooling solvent mixtures for solvation free energy predictions. Chemical Engineering Journal 513:162232 (2025). https://doi.org/10.1016/j.cej.2025.162232
10. Karia T, Chaparro G, Chachuat B, Adjiman CS. Classifier surrogates to ensure phase stability in optimisation-based design of solvent mixtures. Digital Chemical Engineering 14:100200 (2025). https://doi.org/10.1016/j.dche.2024.100200
11. Buzzi, S., Vermeire, F., Al Ibrahim, E., & Green, W. (2025, September). Predicting the Solubility of Organic Compounds in Mixtures with Thermodynamic Cycles and Machine Learning. In 8th RSC-CICAG/RSC-BMCS Artificial Intelligence in Chemistry 2025, Date: 2025/09/22-2025/09/24, Location: Cambridge (UK).
12. Frazier, P. I. (2018). A tutorial on Bayesian optimization. arXiv preprint arXiv:1807.02811.
13. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M, Palmer A, Settels V, Jaakkola T, Jensen K, Barzilay R. Analyzing learned molecular representations for property prediction. J. Chem. Inf. Model. 59:3370-3388 (2019). https://doi.org/10.1021/acs.jcim.9b00237
14. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Message passing neural networks. Lecture Notes in Physics :199-214 (2020). https://doi.org/10.1007/978-3-030-40245-7_10
15. Al Ibrahim E, Morgan N, Müller S, Motati S, Green

WH. Accurately predicting solubility curves via a thermodynamic cycle, machine learning, and solvent ensembles. J. Am. Chem. Soc. 147:45057-45069 (2025). https://doi.org/10.1021/jacs.5c13746
16. Malikov, D., Krasnov, L., Kiseleva, M., Meshcheriakova, E., Kuznetsov, F., Elistratov, V., ... & Bezzubov, S. (2025). MixtureSolDB, dataset of solubility values for organic compounds in binary mixtures of solvents at various temperatures.
17. Devotta, S., & Pendyala, V. R. (1992). Modified Joback group contribution method for normal boiling point of aliphatic halogenated compounds. Industrial & engineering chemistry research, 31(8), 2042-2046.