



HAL
open science

Multi-objective optimization strategy for green solvent design via a deep generative model learned from pre-set molecule pairs

Jun Zhang, Qin Wang, Huaqiang Wen, Vincent Gerbaud, Saimeng Jin,
Weifeng Shen

► To cite this version:

Jun Zhang, Qin Wang, Huaqiang Wen, Vincent Gerbaud, Saimeng Jin, et al.. Multi-objective optimization strategy for green solvent design via a deep generative model learned from pre-set molecule pairs. *Green Chemistry*, 2024, 26 (1), pp.412-427. 10.1039/D3GC04354A . hal-04334474

HAL Id: hal-04334474

<https://hal.science/hal-04334474v1>

Submitted on 11 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Multi-objective Optimization Strategy for Green Solvent Design via Deep Generative Model Learned from Pre-set Molecule Pairs

Jun Zhang^a, Qin Wang^b, Huaqiang Wen^a, Vincent Gerbaud^c, Saimeng Jin^a, and
Weifeng Shen^{a,*}

*^aSchool of Chemistry and Chemical Engineering, Chongqing University, Chongqing
400044, China*

*^bSchool of Chemistry and Chemical Engineering, Chongqing University of Science
and Technology, Chongqing 401331, China*

*^cLaboratoire de Génie Chimique, Université de Toulouse, CNRS, INP, UPS, Toulouse,
France*

***Corresponding author.** E-mail: wangq356@mail2.sysu.edu.cn (Qin Wang), and
shenweifeng@cqu.edu.cn (Weifeng Shen)

Abstracts: Green solvent design is usually a multi-objective optimization problem requiring identifying a set of solvent molecules to balance multiple, often trade-off, properties. At the same time, solvents need to address process constraints since the solvent properties impact the process feasibility like in the extractive distillation separation process. Hence, a green solvent multi-objective optimization framework is proposed with EH&S properties, process constraints, and energy consumption analysis, where the molecular design optimization model relies upon the ability of the proposed infinite dilution activity coefficient (IDAC) direct prediction model to accurately predict process properties in addition to molecular properties. The process properties are short-cut properties of the extractive distillation process, namely selectivity and solution capacity. To this end, the proposed IDAC direct prediction

model is employed to prepare molecule pairs with selectivity and solution capacity improvement constraints to train the molecular multi-objective optimization model, which can learn the optimization path from the pre-set molecule pairs and then optimize a given solvent *via* the prediction of a disconnection site and the molecular fragment addition or removal at that site. An extractive distillation process to separate a cyclohexane/benzene mixture is taken as an example to demonstrate the proposed framework. As a result, three candidate green solvents are optimized and designed to recover benzene from the mixtures of benzene and cyclohexane. The proposed green solvent multi-objective optimization framework is flexible enough to be employed in other chemical separation processes, where solvent properties assessment is needed to evaluate the feasibility and performance of the processes.

Keywords: Molecular Multi-objective Optimization; QSPR; Green Solvent Design; Extractive Distillation

1. Introduction

In many separation processes, like azeotropic or extractive distillation or liquid-liquid extraction, the solvent is needed to perform the desired separation. Solvent design is inherently a constrained multi-objective optimization problem.^{1, 2} The first set of constraints concerns matching desired solvent property values. These properties are multiple and usually cover not only molecular properties, related to the primary function of the solvent, like solubilizing an active principle or having a preferred affinity with one of the molecules in a mixture, but also other properties that may ease the process operation. Model-based property predictions are numerous but are confronted with various challenges, like coping with stereoisomers for group-

contribution methods, or sampling correctly the vast solvent search space spanning the chemistry field for computer-costly quantum mechanical methods.

Besides in separation processes, the process feasibility sets additional constraints on the solvent. Hence, a search simultaneously combining molecular and process constraints is a challenge which is the purpose of our contribution, and which would be facilitated by using model-based approaches to optimize the structure of solvent molecules. But the successive optimal solvent design first followed by an optimal process design bears a risk of error propagation that could rule out the whole procedure.

In this case, we proposed a molecular multi-objective optimization model to purposefully modify the structure of solvent molecules with some drawbacks (such as EH&S negative impacts) to obtain the green solvent with the desired separating performance rather than simply utilizing a molecular generative model to enlarge the chemical space for subsequent solvent screening with multi-index constraints. The multi-objective optimization model can learn the optimization path from the pre-set molecule pairs. Every two paired molecules (M_x , M_y) in the pre-set molecule pairs are similar in their molecule structures and only have a single different disconnection site, but the scores of both selectivity and solution capacity of M_y are at least 20% larger than those of M_x . The prepared pre-set molecular pairs are used to train the proposed molecular multi-objective optimization model, which can learn the difference between the molecular pairs and can learn the optimization path from M_x to M_y . To prepare the molecule pairs, an improved deep learning-based IDAC direct prediction model trained over a COSMO-SAC database is developed for predicting the selectivity and

solution capacity of the molecule pairs. The proposed IDAC direct prediction method can provide superior predictive performance compared with the IDAC indirect prediction method, which first predicts the V_{COSMO} and 51 σ -profile and then calculates IDAC by the COSMO-SAC model. The indirect IDAC prediction method could result in more information lost during the prediction and COSMO-SAC calculation process. The improved deep learning-based direct IDAC prediction model is integrated with the molecular multi-objective optimization model to form the proposed green solvent multi-objective and multi-scale optimization framework with EH&S properties and process constraints, and energy and economic analysis. The proposed green solvent multi-objective and multi-scale optimization framework can:

- 1) Simultaneously optimize multiple trade-off properties such as the selectivity and solution capacity of the solvent;
- 2) Learn from the pre-set molecule pairs that have similar molecular structures but have differences in the interested properties;
- 3) Visualize the optimization path of the solvent molecular structure;
- 4) Accurately and directly predict the IDAC of the molecules.

The paper is organized as follows: the next section 2 gives a non-exhaustive overview of solvent design issues, related computer-aided approaches, and connections to some process design issues for extractive distillation processes. Section 3 describes the integrated molecular multi-objective and multi-scale optimization framework. Section 4 describes and evaluates the performance of the improved model for the direct prediction of the infinite dilution activity coefficient using deep learning techniques. Section 5 introduces the molecular multi-objective optimization model. Section 6 is an illustrative case study about solvent optimization

and design for an extractive distillation process.

2. Background

When designing a solvent, one should match target values for properties that directly impact the process separation feasibility like the selectivity and solution capacity in liquid-liquid separation, melting point in solid separation processes, *etc.* At the same time, properties that affect the process performance and operation, in terms of economics and energy requirements, should also be considered, like boiling point for the distillation process, molar volume for batch processes along with properties related to transport phenomena like viscosity, surface tension, heat capacity, *etc.*³ Nowadays, the sustainability of new solvents (e.g., toxicity, safety, environmental impact, *etc.*)⁴⁻⁶ is also becoming a key design objective,⁷ especially important for the green solvent design task, and for complying with regulations such as US Toxicity Characteristic Leaching Procedure (TCLP) or EU Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH) ones that affect process authorizations. Considering the vast number of potential solvents, the trial-and-error method for solvent identification may be highly time-consuming and even unrealistic when one considers only a single property. In addition, promising solvents could be missed if the trial-and-error method relies on a fixed solvent database. Hence, model-based solvent selection or design methods, like computer-aided molecular design (CAMD), are extremely desirable to address the issue of exploring the vast solvent search space.⁸

In support of any computer-aided solvent design approach, one needs to access solvent property values. Evaluation of the solvent thermodynamic properties requires

measuring them or calculating them by using property estimation models because they are more appropriate in a preliminary process design phase. Hence, property calculation or estimation models play a significantly important role in model-based solvent design methods since they can correlate molecular structures with solvent thermodynamic properties. For any property of interest in a real process, there exists a variety of property models, and choosing the most suitable models is a key step.⁹ Each model bears different accuracies, predictive capabilities, and computation costs.¹⁰ The property estimation methods mainly include descriptor-based methods,¹¹ group-contribution (GC) methods,¹² quantum mechanical (QM) methods,¹³ and deep learning (DL) methods.^{14, 15} For example, for extractive distillation, the process we select for illustration, one real property of relevance might be stated as solubilizing an active principle or having a preferential affinity with one of the compounds in the mixture to be separated. It can be evaluated by various models, either by comparing the similarity of Hansen solubility parameter values between solvent and molecule of interest, a simple correlative model with no access to temperature dependency; or by solving thermodynamic phase equilibrium for computing solubility with temperature dependency; or by comparing interaction surface potential like COSMO sigma potential curves, which requires quantum mechanics calculations. The GC method is one of the most widely utilized and efficient techniques to evaluate macroscopic physicochemical properties. However, the performance of first-order GC models, with contributions regressed over experimental data directly related to the occurrence of simple chemical groups like -CH₂, -OH, *etc.*, is sometimes weakened because they cannot take account of the proximity effects and distinguish between isomers.^{3, 16} To

address these issues, second and third-order GC models are developed for discriminating the structural isomers.¹⁷ but they are still deficient for many stereoisomers such as *cis/trans* isomers.¹⁸ These issues can be tackled with quantum mechanical-based (QM-based) solvation models, such as COSMO-RS^{19, 20} and COSMO-SAC^{21, 22}. With just a few parameters such as the surface charge density profile (σ -profile) and the cavity volumes (V_{COSMO}), the COSMO-based models can achieve a decent accuracy for the calculation of thermodynamic properties. However, the initial QM calculations bear a heavy computer cost and are highly time-consuming, and even unrealistic when exploring the vast search space of solvent molecules.²³ To this end, GC-COSMO techniques were proposed as a shortcut to more efficiently access the V_{COSMO} and σ -profile.^{24, 25} However, due to the inherent GC limitations, these GC-COSMO techniques not only have difficulties in appropriately handling isomers and proximity effects but also limited in the variety of functional groups available in open-source databases. With the availability of the COSMO-type databases (e.g., the VT-2005²⁶), as an alternative, deep learning-based (DL-based) techniques²⁷⁻³⁰ can be applied as another shortcut to obtain the σ -profile and V_{COSMO} .^{31, 32} However, the VT-2005 database only contains 1431 compounds, which could be not enough to train a DL-based prediction model with satisfying generalization ability. Additionally, such DL-based prediction models are developed to predict the V_{COSMO} and the σ -profile, and then using the predicted parameters to calculate the IDAC. This indirect IDAC calculation process could lead to a decline in accuracy.

Once property estimation models are available, computer-aided molecular design³³ (CAMD) is an effective approach for screening existing solvents and

designing new ones. In CAMD, a collection of pre-prepared molecular functional groups are assembled to generate potential solvents through mixed integrated linear programming (MILP) or mixed integrated non-linear programming (MINLP) or stochastic algorithms with the objective functions and constraints (such as molecular structural, property, and process operating constraints).³⁴⁻³⁷ However, with the increase in the number of preselected functional groups, the CAMD method may face the problem of combination explosion.^{3, 38}

Recent advancements in the domain of artificial intelligence have accelerated the development and application of techniques for inverse molecular design.³⁹⁻⁴² For instance, molecular generation models have been applied in many fields.^{38, 43} Molecular graph generation techniques⁴⁴ as an outstanding representative have become one of the most widely adopted approaches for molecule design. Recently, a fragment-based hierarchical encoder-decoder model for molecular generation was proposed by Jin et al.⁴⁵ The fragments extracted from the training molecules are analogous to the molecular functional groups used in the group contribution methods. The molecular fragments could integrate knowledge from the chemistry domain interpretability into the model.⁴⁶ Molecules can be optimized by predicting a disconnection site and performing the molecular fragment addition or removal at that site. However, this model cannot simultaneously optimize multiple trade-off properties of solvent molecules. Therefore, this kind of single objective optimization model is very difficult to couple with the multi-dimensional and highly nonlinear chemical separation process. Although there are deep molecular optimization models labeled “multi-objective”.^{47, 48} These models usually aggregate multiple objectives

into a single scalar objective.

However, solvent property knowledge is only a first step in the design of a performant separation process, of which the process model can be highly nonlinear because the process feasibility is often directly related to the solvent characteristics. For example, there are some trade-off properties such as the selectivity and solution capacity that are not perfectly correlated and, therefore, molecular multi-objective optimization cannot be addressed by these models. Hence, some authors have explored the simultaneous design of the solvent and the process attributes in a so-called reverse engineering Computer Aided Molecule and Process Design (CAMPD) approach. For example, some authors have proposed a framework for the integrated design of solvent and extractive distillation process by solving a multi-objective optimization problem addressing constraints related to thermodynamic process feasibility, along with process operation, process model, and molecular constraints,⁴⁹ or more rigorous rate-based model.⁵⁰ In these works, the property prediction in molecular scale is addressed using COSMO approaches while the process model can be a pinch-based model based on minimum solvent flow rate and minimum energy demand⁴⁹ or a more rigorous rate-based model.⁵⁰ The use of such process models is relevant for an accurate process design but there exist simpler criteria for assessing extractive distillation feasibility, like solvent capacity and selectivity,⁵¹ which are further related to infinite dilution activity coefficients (IDAC), and univolatility curves.⁵² In this contribution, we propose a molecular multi-objective and multi-scale optimization framework for the combined molecular and process design with the predicted process constraints (solvent selectivity and capacity based on IDAC) where

the process-related properties are directly used to train the molecular structure optimization model, with the help of deep-learning techniques.

3. The deep learning-based molecular multi-objective and multi-scale optimization framework for green solvent design

The deep learning-based molecular multi-objective and multi-scale optimization framework for green extraction distillation solvent design is presented by integrating an improved deep learning-based model for IDAC direct prediction (in **Section 4**) and a data-driven deep molecular multi-objective optimization model (in **Section 5**) as shown in Figure 1. Meanwhile, the EH&S property constraints, process constraints, and energy consumption analysis are considered to ensure the sustainability and technological economy of green solvents.

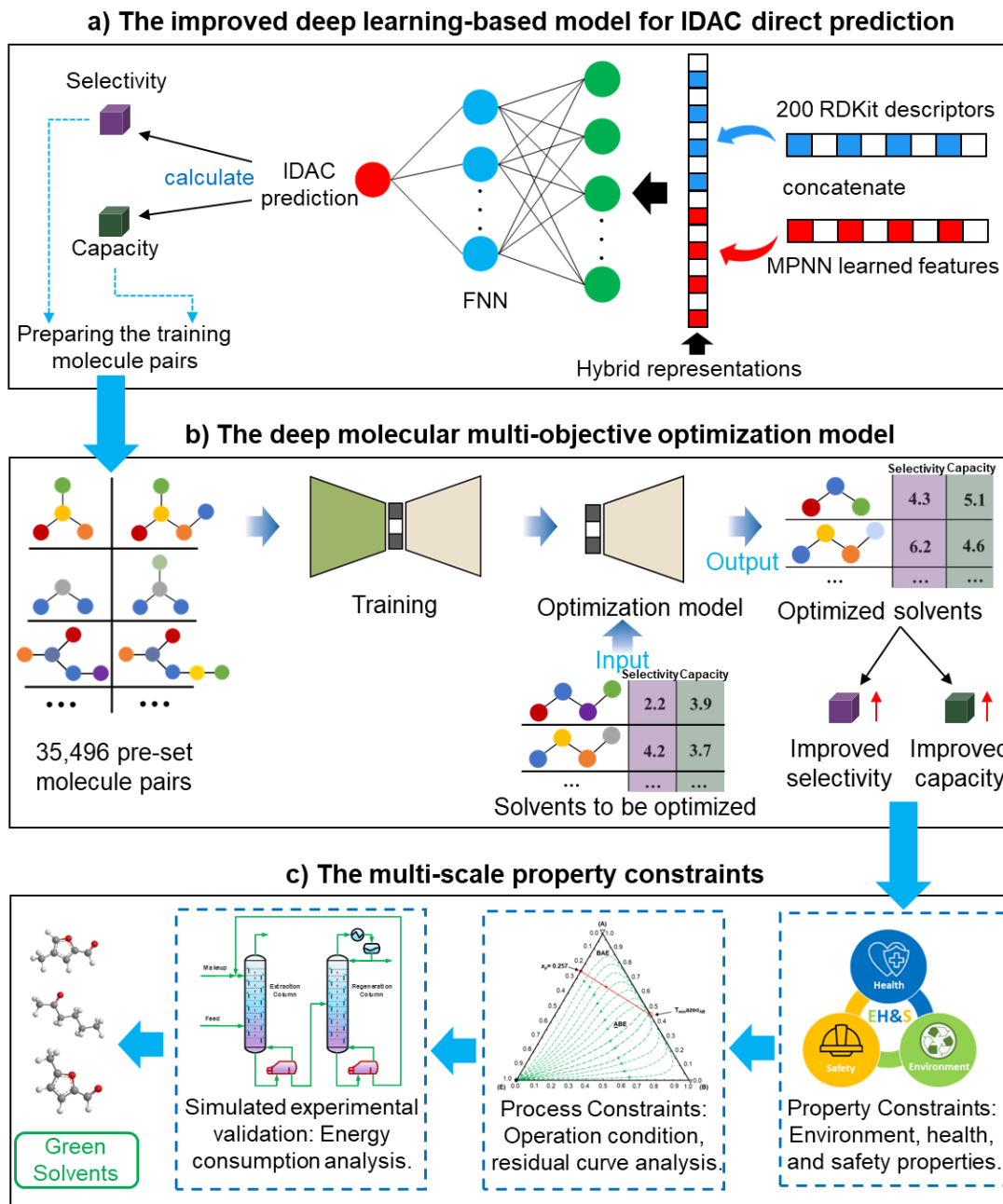


Figure 1. The green solvent multi-objective and multi-scale optimization framework towards extractive distillation processes.

The proposed framework for green solvent design will be applied to an extractive distillation process to separate cyclohexane and benzene mixtures (in **Section 6**).

4. An improved deep learning model for the infinite dilution activity coefficient direct prediction

4.1 Data preparation

The COSMO-SAC model²² is utilized in this contribution for the IDAC calculation. The UD database⁵³ contains quantum mechanically derived V_{COSMO} and σ -profile for 2,261 compounds. Heuristically, the increase in molecular weight of a solvent results in a higher normal boiling point, which usually means a higher energy consumption for an extractive distillation process, and reduces its economic viability. Therefore, only the molecules with less than 12 root atoms (hydrogen atoms ignored) are considered in this contribution. Additionally, as a collected dataset, data cleaning is essential to be applied to remove outliers. The Pauta criterion,⁵⁴ also known as the three Sigma rule, is employed for the data cleaning process. After the data preparation process, 2,130 compounds are remaining in the UD database. The quantum mechanically derived V_{COSMO} and σ -profile of the 2,130 compounds (2,125 compounds for model training and 5 compounds for external test) are provided in Table S1 in the Supplementary Material. The calculated IDAC of the 2,130 compounds in cyclohexane (A) and benzene (B) is detailed in Table S2 in the Supplementary Material.

4.2 Development of the deep learning model for IDAC direct prediction

There are two paths to calculate the IDAC of a molecule in a certain solvent: 1) The

V_{COSMO} and 51 σ -profile predictive models are trained, and then the IDACs of different compounds in a certain solvent are calculated utilizing the COSMO-SAC with the estimated parameters as shown in Figure 2a. This indirect IDAC calculation process can be termed as indirect method (IM) in this work; 2) The IDACs of different compounds in a certain solvent are directly calculated employing the COSMO-SAC, and then the calculated IDAC information is utilized to train an IDAC predictive model as illustrated in Figure 2b. This direct IDAC calculation path can be termed as direct method (DM). The IM-based IDAC calculation path has been introduced in our previous work.^{32, 55} In this contribution, the DM-based IDAC predictive model is developed to evaluate which IDAC calculation path performs better.

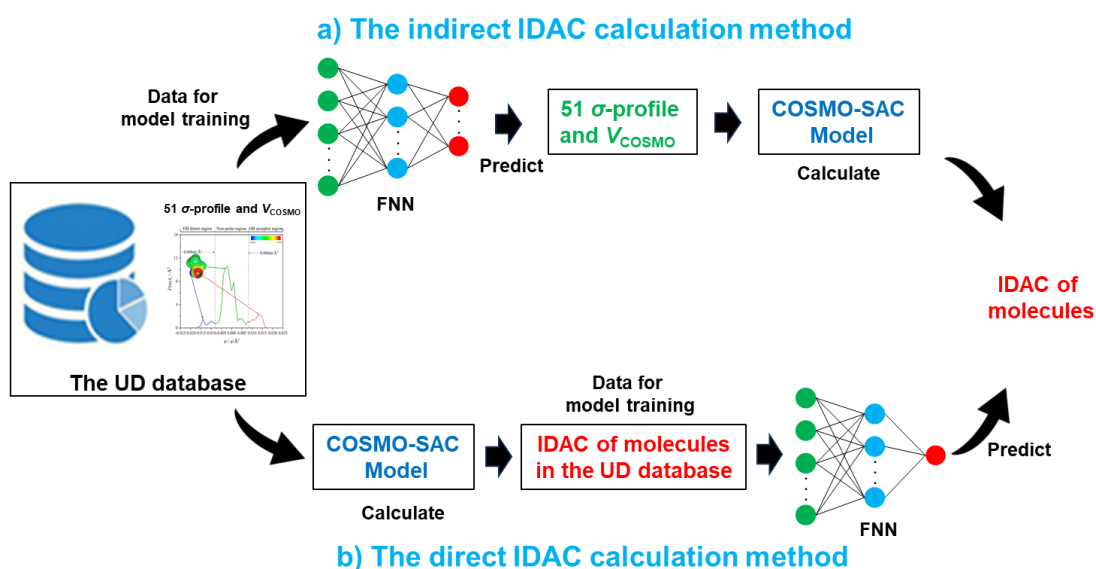


Figure 2. The schematic diagram for IDAC calculation. a) an indirect method (IM). b) a direct method (DM).

First, the IDACs of the 2,130 compounds in benzene and cyclohexane are calculated using the COSMO-SAC model with their V_{COSMO} and σ -profile information from the UD database. Subsequently, the hybrid representations^{28, 32} of the 2,125

compounds (another five compounds are used as the external validation data) are utilized as input to train the feedforward neural network for the IDAC prediction in benzene and cyclohexane (IDAC-benzene and IDAC-cyclohexane) as shown in Figure 3. The message-passing neural network (MPNN) is a graph neural network, which consists of two phases, namely, the message-passing phase and the readout phase²⁸. In the message-passing phase, the MPNN updates information on the directed bonds, as shown in Figure 3. In the readout phase, a readout function is utilized to provide a vector representation of the molecular structure. The MPNN learned descriptors mainly focus on the local information about molecular structure due to the message updating mechanism. Therefore, the molecule level 200 dimensional RDKit calculated descriptors (as shown in Figure 3) that can capture the global information of the molecular structure are employed to integrate with the MPNN learned features to form the molecular hybrid representation, which can retain the molecular local and global information as much as possible. The data split setting for training the two proposed models is 0.8:0.1:0.1. The early stopping technique is employed to avoid overfitting. Finally, the 10-fold cross-validation (10-fold CV) method is applied to improve the stability of the proposed two models. In this contribution, the hidden size of MPNN, the depth of MPNN, the layer number of FNN, and the dropout of FNN are optimized with the Bayesian Optimization method embedded in a Python package hyperopt.⁵⁶

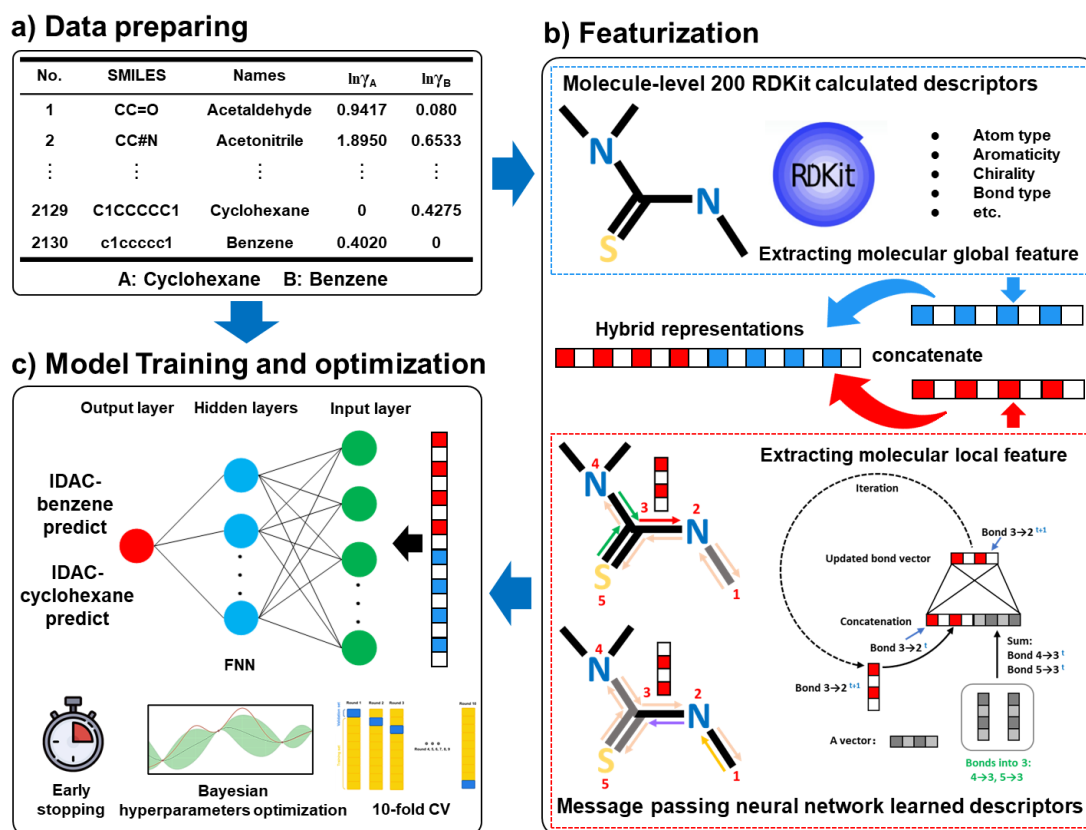


Figure 3. The scheme diagram of the proposed IDAC direct method (DM) predictive models.

4.3 The performance evaluation of the proposed deep learning prediction models

The optimal hyperparameter combinations for the proposed IDAC-benzene and IDAC-cyclohexane predictive models are summarized in Table 1.

Table 1. The optimal hyperparameter combinations for the proposed prediction models of the IDAC-benzene and IDAC-cyclohexane.

Hyperparameters	Range	IDAC-benzene	IDAC-cyclohexane
Hidden size	[300,3000]	1200	1300
Depth	[2,7]	6	6
Dropout	[0,0.4]	0.0	0.0
Number of layers	[1,5]	3	3

In this contribution, three evaluating metrics, *i. e.* the mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R^2), are adopted as the evaluation criteria. The prediction performance of the IM and DM with the UD database is summarized in Table 2. In addition to the FNN model, the prediction

performance using random-forest and support-vector machine approaches is also summarized in Table 2 to explore which machine learning approach is more suitable for IDAC prediction. The optimal hyperparameter combinations of the random forest and support vector machine-based approaches are detailed in Table S3. Based on the statistical analysis, the FNN-based models (IM and DM models) have superior predictive performance over the random forest and support vector machine-based models. The performance of the 10-fold CV of the proposed DM models for the IDACs in benzene and cyclohexane prediction on the test sets are both better than those of the IM predictive model.

Table 2. The 10-fold cross-validation performance of the indirect method (IM) and direct method (DM) predictive models.

(a) IM³² (FNN-based model)			
	10 CV MAE	10 CV MSE	10 CV R^2
IDAC-benzene	0.1216 ± 0.0140	0.0720 ± 0.0163	0.8651 ± 0.0338
IDAC-cyclohexane	0.1755 ± 0.0180	0.1435 ± 0.0341	0.9123 ± 0.0198
(b) DM (FNN-based model)			
	10 CV MAE	10 CV MSE	10 CV R^2
IDAC-benzene	0.1146 ± 0.0108	0.0506 ± 0.0084	0.9036 ± 0.0128
IDAC-cyclohexane	0.1652 ± 0.0173	0.1126 ± 0.0262	0.9257 ± 0.0226
(c) Random forest-based model			
	10 CV MAE	10 CV MSE	10 CV R^2
IDAC-benzene	0.2224 ± 0.0198	0.1581 ± 0.0221	0.6985 ± 0.0314
IDAC-cyclohexane	0.3381 ± 0.0264	0.3394 ± 0.0689	0.7814 ± 0.0367
(d) Support vector machine-based model			
	10 CV MAE	10 CV MSE	10 CV R^2
IDAC-benzene	0.2516 ± 0.0164	0.1456 ± 0.0206	0.7213 ± 0.0389
IDAC-cyclohexane	0.3571 ± 0.0263	0.2965 ± 0.0476	0.8089 ± 0.0196

In addition to the above-mentioned statistical analysis, five molecules in the external validation dataset are utilized as examples to evaluate the ability of the proposed predictive models to discriminate the stereoisomers and structural isomers and to deal with complex molecules. In this work, the external validation dataset consists of *N,N*-Diethylaniline (complex compounds), *P*-xylene and *O*-xylene (structural isomers), and *Cis*-3-hexene and *Trans*-3-hexene (*cis/trans* isomers). The

heteroatomic nitrogen in *N,N*-Diethylaniline has an inducing effect on the delocalized π electron system of the aromatic ring, which could lead to a poor prediction performance by some Quantitative Structure-Property Relationships (QSPR) models³¹. O-xylene and P-xylene are a pair of structural isomers. *Trans*-3-hexene and *Cis*-3-hexene are a pair of *cis/trans* isomers. The predictive performance of IM and DM models is tabulated in Table 3. Regarding the *N,N*-Diethylaniline, the proposed DM models can achieve a better predictive performance than IM models. Regarding the structural and *cis/trans* isomers, both IM and DM models have a satisfactory ability to differentiate isomers. Additionally, we visualized the chemical space of the training and external dataset by projecting the Morgan fingerprints (radius=2, 1024 bits) of the molecules onto the 2D space (as shown in Figure 4) *via* the t-SNE approach⁵⁷. As shown in Figure 4, the five external data are not very similar to the well-represented molecules in the training dataset. Moreover, the five external data are scattered in different regions of the chemistry space of the training dataset. Therefore, the proposed DM models have decent IDAC predictive performance and good generalization ability.

Table 3. The prediction performance of the indirect method (IM) and direct method (DM) models on the external validation dataset.

Compounds names	IDAC-cyclohexane			IDAC-benzene		
	QM derived values	IM	DM	QM derived values	IM	DM
<i>N,N</i> -Diethylaniline	0.3215	0.2398	0.3028	-0.1047	-0.0745	-0.1033
P-xylene	0.2230	0.2194	0.2295	0.0008	0.0109	-0.0004
O-xylene	0.2654	0.2683	0.2472	0.0007	-0.0037	0.0014
<i>Cis</i> -3-hexene	0.0550	0.0505	0.0463	0.1860	0.1847	0.2072
<i>Trans</i> -3-hexene	0.0457	0.0475	0.0505	0.2325	0.2165	0.2183

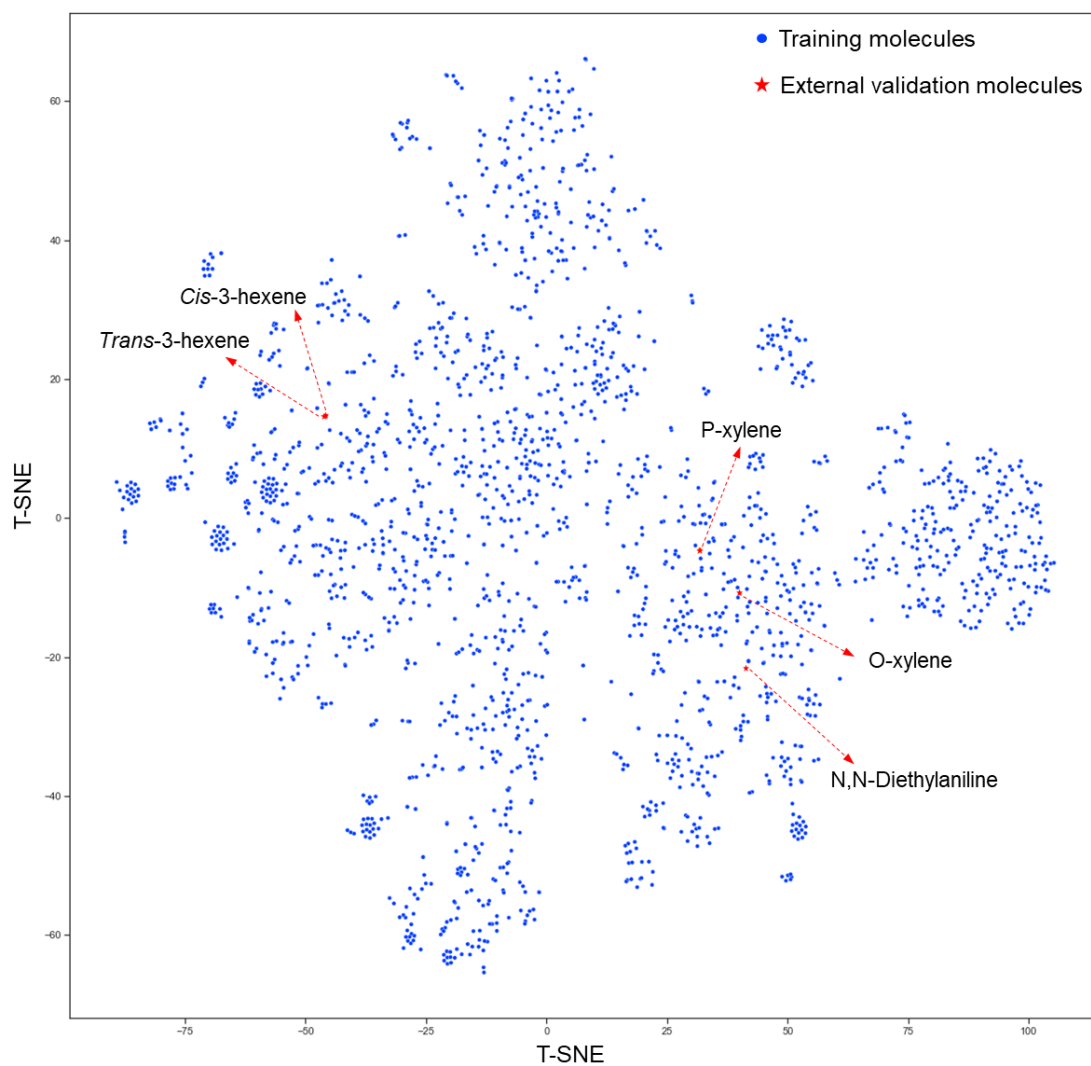


Figure 4. The chemical space of the training and external dataset visualized *via* the t-SNE approach.

Based on the predictive performance analysis mentioned above, the proposed DM models have a better generalization ability than the IM models. Additionally, the proposed IDAC prediction models can discriminate the isomers, including the isomers and *cis/trans* isomers, and can deal with complex compounds such as hetero-atom compounds.

5. An interpretable molecular multi-objective optimization model learned from pre-set molecule pairs

5.1 Training data preparing for molecular multi-objective optimization

The ChEMBL dataset⁵⁸ processed by Olivecrona⁵⁹ is used to construct the molecule pairs, which are employed as the training data for solvent molecular multi-objective optimization. There are 1,179,477 compounds in the processed ChEMBL dataset. Each compound is restricted to contain 10 to 50 root atoms and only has atoms in {H, B, C, N, O, F, Si, P, S, Cl, Br, I}. The training molecule pairs are constructed as follows. First, there are 18,155 molecules are identified from the processed ChEMBL dataset with root atoms not more than 12. We adopt the 12 root atoms threshold because larger molecules usually have a higher normal boiling point, and a molecule with a high normal boiling point is not suitable to use as a solvent to separate the benzene and cyclohexane via extractive distillation. Second, there are $C_{18,155}^2 = 164,792,935$ molecule pairs (M_x, M_y) constructed from 18,155 processed molecules. Third, 1,590,350 molecule pairs have similarities $\text{sim}(M_x, M_y) \geq 0.4$. The similarities of the molecule pairs can be measured by the Tanimoto coefficient over 2,048-dimension binary Morgan fingerprints with radius 1. The similarity threshold is adopted because the proposed molecular optimization model needs the training molecule pairs with only one fragment different at one disconnection site, which can improve the learning efficiency of the molecular optimization model. Fourth, the DF-GED algorithm is used to extract molecule pairs that have only one fragment different at one disconnection site, which can improve the learning efficiency of the molecular multi-objective optimization model. There are 100,629 molecule pairs extracted from

the 1,590,350 pairs of molecules. Fifth, among the 100,629 pairs of molecules, we select the molecule pairs that meet the following property constraints: for the selectivity, the selectivity score of M_y should be improved by at least 20% compared with M_x in a molecule pair, that is,

$$\frac{S(M_y)-S(M_x)}{S(M_x)} \geq 0.2; \quad (1)$$

for the solution capacity score of M_y should be also improved by at least 20% compared with M_x in a molecule pair, that is,

$$\frac{C(M_y)-C(M_x)}{C(M_x)} \geq 0.2; \quad (2)$$

As a result, 35,496 molecule pairs (detailed in Table S4 in the Supplementary Material) are identified that can be used as the training data with the property constraints.

5.2 Development of the multi-objective optimization model of solvent molecules

The proposed multi-objective molecular optimization approach of solvent molecules extends the hierarchical generation model to multi-objective molecular optimization by learning the molecule pairs with improved selectivity and solution capacity.

In the hierarchical encoding process, a molecule can be represented by a hierarchical graph with three layers⁴⁵, i.e., atom layer, attachment layer, and fragment layer, as seen in Figure 5a. The details of the fragment extraction approach and the hierarchical encoding method are introduced in the contribution presented by Jin et al.⁴⁵ and Chen et al.⁴⁶ In the hierarchical molecular representation framework, a molecule graph Γ can be represented as a set of fragments Φ_1, \dots, Φ_n , and their attachments A_1, \dots, A_n . Each attachment A_i in this layer denotes a specific attachment configuration of fragment Φ_i , including the connection information between Φ_i and

one of its neighbor fragments. In the atom layer, a molecule can be depicted as graph $\Gamma_x = (A_x, B_x)$, where A_x and B_x represent the atoms and corresponding bonds in M_x . In the attachment layer, molecule Γ_x is constituted by a series of fragments Φ_1, \dots, Φ_n extracted from the M_x . In the fragment layer, a molecule M_x is represented as a tree-constructed graph T_x . The tree-constructed representation can be depicted as $T_x = (\zeta_x, E_x)$,⁴⁴ where all the fragments in M_x are extracted as nodes in ζ_x ; nodes with the same atoms are connected with edges in E_x . The encoder encodes the molecule pairs (M_x, M_y) as graph (Γ_x, Γ_y) using message passing networks, and as a tree-constructed graph (T_x, T_y) using tree message passing networks.

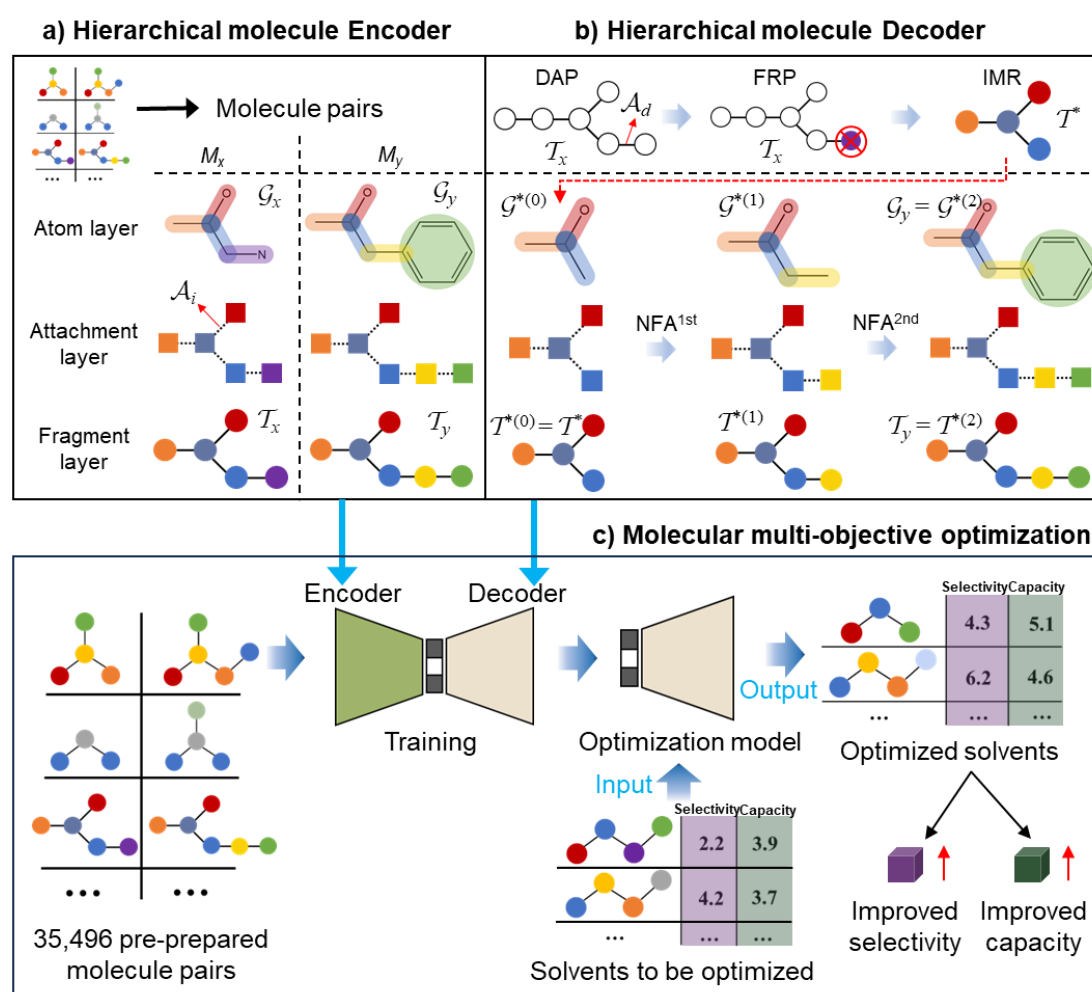


Figure 5. The schematic diagram of the hierarchical molecule (a) encoder, (b) decoder,

and (c) multi-objective optimization process.

In the hierarchical decoding process, the decoder conducts a series of modified operations that optimize M_x into M_y , as seen in Figure 5b. The details of the hierarchical decoding method are introduced in the work presented by Jin et al⁴⁵ and Chen et al.⁴⁶ First, the decoder performs disconnection attachment prediction (DAP) to find an attachment A_d in T_x as the disconnection site. Second, at neighbors of A_d , the decoder performs fragment-removing prediction (FRP) to remove fragments attached to A_d . Third, an intermediate representation (IMR) for the remaining scaffold ($\Gamma^{*(0)}$, $T^{*(0)}$) is produced after the fragment removal operation. Fourth, over ($\Gamma^{*(0)}$, $T^{*(0)}$), the decoder conducts new fragment attachment (NFA) prediction iteratively to optimize M_x into M_y . The optimal graph edit paths can be identified by the DF-GED algorithm⁶⁰.

By learning from the selectivity and solution capacity improved molecule pairs (training molecule pairs), the hierarchical molecular multi-objective optimization model can realize the multi-objective optimization of solvent molecules as illustrated in Figure 5c.

6. Case study of the green solvent multi-objective and multi-scale optimization framework






A case study using extraction distillation to separate aliphatic and aromatic mixtures⁶¹ is used to evaluate the proposed green solvent multi-objective optimization framework. In this work, the mixtures of aromatic/aliphatic are simplified as the mixtures of cyclohexane (A)/benzene (B).⁶² The green extractive distillation solvent multi-objective optimization framework can be decomposed into three steps, *i.e.*, molecular multi-objective optimization, properties constraints, and process constraints

as introduced in Section 2.

6.1 The extractive distillation solvent multi-objective optimization

In this step, as the inputs of the molecular multi-objective optimization model, industrial extractive solvent molecules that need to be optimized should be first identified. Five widely employed extractive distillation solvents for separating the mixtures of cyclohexane (A) and benzene (B) are summarized in Table 4 based on extensive literature research.^{61, 63-66} However, all these solvents have some drawbacks, such as toxicity or ecological hazard. The toxicity and ecological information of these solvents (with experimentally measured or predicted properties) are available in the Syntelly database.⁶⁷

Table 4. Five commonly utilized extractive distillation solvents for separating benzene from the mixtures of benzene/cyclohexane as inputs of the molecular multi-objective optimization model.

Names	Structure	Drawbacks	References
Furfural		Toxicity	61
Sulfolane		Ecological hazard	63
DMSO		Toxicity	64
DMF		Ecological hazard	65
NMP		Ecological hazard	66

Taking the five common industrial solvent molecules as inputs of the molecular multi-objective optimization model, 20 optimized solvent molecules are generated for every single widely used solvent (as seen in Figures 4a, b, c, d, and e) *via* the trained molecular multi-objective optimization model introduced in Section 4. Accordingly, 100 optimized solvent molecules are generated as tabulated in Table S5 in the Supplementary Material.

6.2 EH&S property constraints

In this step, the 100 optimized solvent molecules are screened by EH&S properties. The environmental property takes into account three ecological indicators, *i.e.*, bioconcentration factor, 40 hours *Tetrahymena pyriformis* IGC₅₀, and 48 hours *Daphnia magna* LC₅₀. If a solvent negatively affects the environment, it will be marked in red in Figure 6. The health properties can be quantified by the dosage of rat oral. The threshold value for toxicity is 2000 mg/kg. If the rat oral of a given solvent is 500 mg/kg, it will negatively affect health and will be marked in red in Figure 6. The safety can be quantified by the flash point. For a given solvent, the higher its flash point, the better for the storage security. In this work, if the flash point is above 280 K,³⁶ it will positively impact the storage security and it will be marked in green color in Figure 6. All the EH&S information can be collected from the Syntelly database.⁶⁷ As a result, 10 solvent molecules remain screened by EH&S properties constraints and are displayed in Table 5.

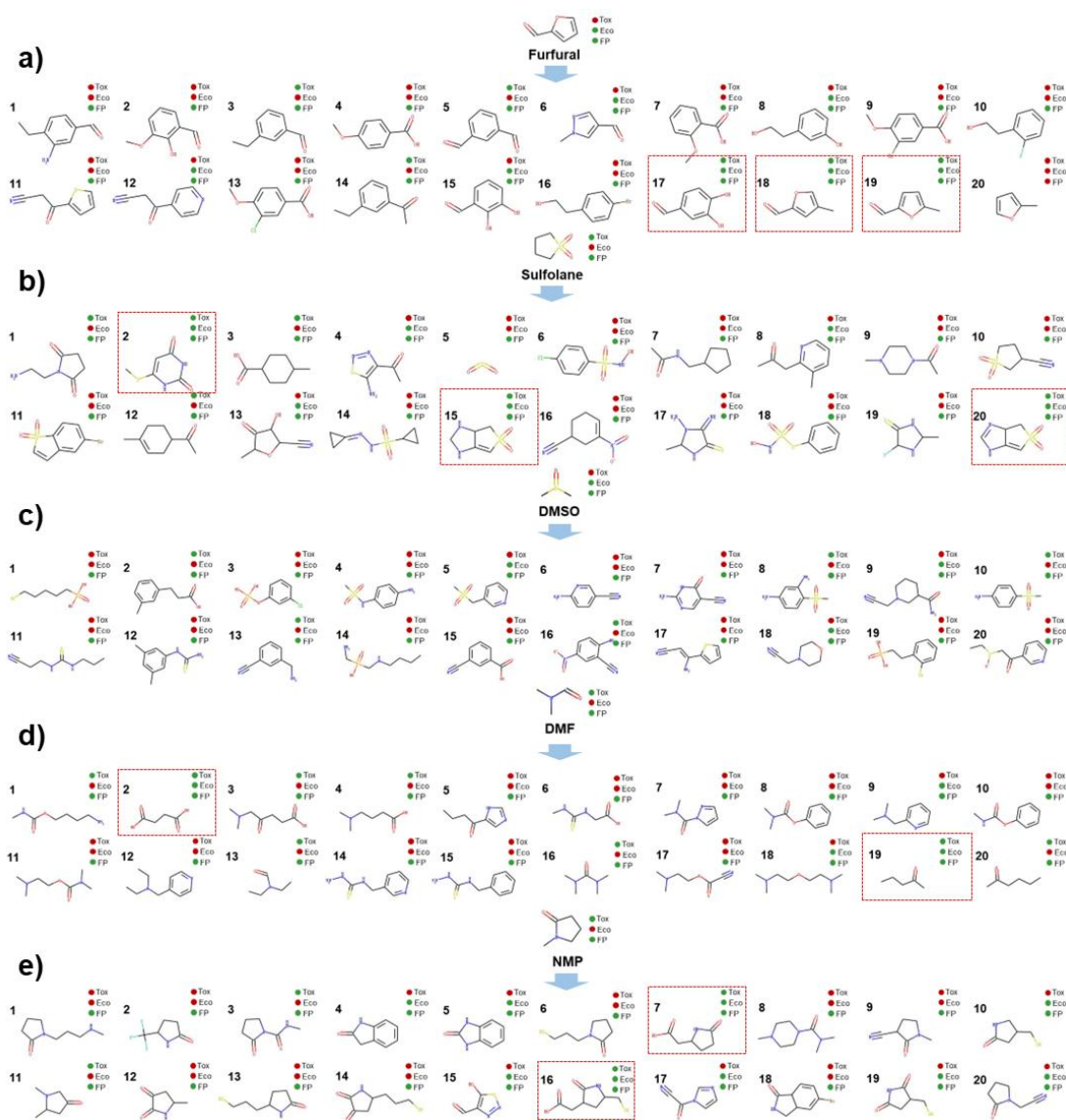


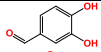
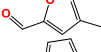
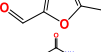
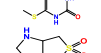
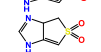
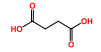
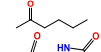
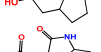
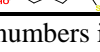
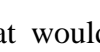
Figure 6. EH&S properties information on the 100 optimized solvent molecules with the proposed molecular multi-objective optimization model. Tox, Eco, and FP are the abbreviations for toxicity, ecology, and flash point. The green color denotes positive (or good) properties for EH&S constraints, and red means negative. The red dotted boxes mark all solvent molecules with three positive EH&S properties.

6.3 Process constraints

In this step, process operation conditions are quantified by normal melting point and normal boiling point. The melting point of the solvent should be below 310 K^{36} to ensure that it is in the liquid state at operating temperature. The boiling point of the

solvent should be below 580 K³⁶ to meet the relatively economical separation energy consumption. There are 3 solvent molecules (*i.e.*, a18, a19, and d19, whose names are 4-methyl furfural, 5-methyl furfural, and 2-hexanone, respectively) remaining after the operation conditions screening. Detailed normal melting point and boiling point information on the 10 solvent molecules after EH&S properties screening is provided in Table 5.

Table 5. The melting point and boiling point information of the 10 solvent molecules after EH&S properties screening.

Names ^a	Smiles	Structure	Melting point/K	Boiling point/K
a17	<chem>O=Cc1ccc(O)c(O)c1</chem>		413.15	550.15
a18	<chem>Cc1coc(C=O)c1</chem>		303.15	455.15
a19	<chem>Cc1ccc(C=O)o1</chem>		293.15	460.15
b2	<chem>CSc1cc(=O)[nH]c(=O)[nH]1</chem>		523.15	578.15
b15	<chem>O=S1(=O)C=C2NCNC2C1</chem>		473.15	564.15
b20	<chem>O=S1(=O)C=C2NC=NC2C1</chem>		473.15	547.15
d2	<chem>O=C(O)CCC(=O)O</chem>		461.15	546.15
d19	<chem>CCCCC(C)=O</chem>		217.65	400.75
e7	<chem>O=C(O)CC1CCC(=O)N1</chem>		430.15	570.15
e16	<chem>O=C(O)CC1CC(CS)NC1=O</chem>		430.15	559.15

a: The names correspond to the serial numbers in Figure 6.

To further screen solvents that would make the extractive distillation process feasible, the residue curve analyses of the 3 screened solvents are conducted as tabulated in Figure 7. According to Gerbaud et al.'s review,⁵² the combined analysis of residual curve maps (RC) and univolatility line can help evaluate whether a solvent is suitable or not for a mixture separation *via* extractive distillation.⁶⁸ As illustrated in the RC maps, every single curve originates from the azeotrope point and terminates in the pure component. Additionally, there is one distillation region for each of the three RC maps. In the residue curve map, A or B are saddle points of the distillation region and cannot be obtained by azeotropic distillation. On the other hand, the univolatility

line splits the ternary diagram into two volatility order regions for all three solvents. With the feeding of the solvent at another location than the main feed, the extractive distillation process enables obtaining the most volatile component in the volatility order regions where the solvent lies.⁵² This is the case for cyclohexane with the 3 green candidate solvents. Therefore, it is possible to separate the benzene/cyclohexane mixtures as pure products, first by removing cyclohexane from the extractive distillation column, then by recovering benzene as a distillate from the regeneration column where high-purity solvent is obtained in the bottom and then recycled to the extraction distillation column. The intersection point x_p of the isovolatility curve with the triangle edge largely determines the minimum usage of the solvent.^{52, 69, 70} The lower the x_p , the less the amount of solvent is required. As we can see, the mole usage amount of 2-hexanone is more than that of 4-methyl furfural and 5-methyl furfural. The results of the combined residue curve map and univolatility analyses can further prove that the proposed IDAC predictive models can achieve reliable and accurate prediction performance.

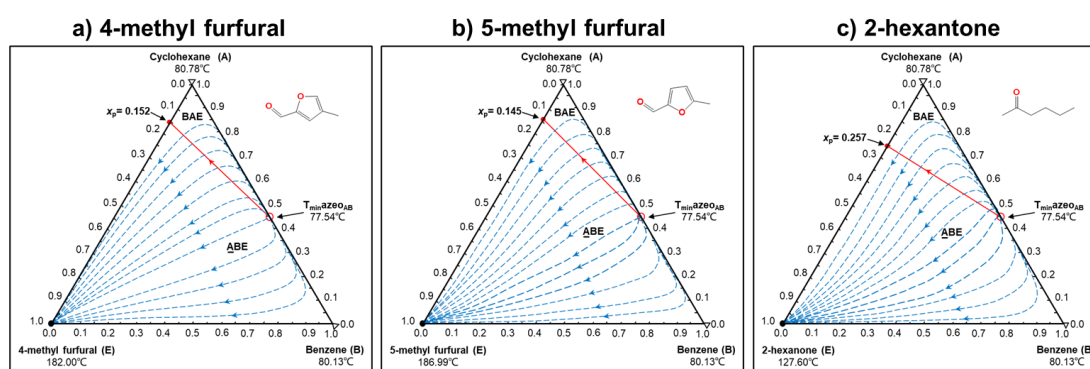










Figure 7. The residue curve maps of (1) 4-methyl furfural, (2) 5-methyl furfural, and (3) 2-hexanone in the cyclohexane (A)/benzene (B) mixtures.

6.4 Energy consumption analysis

The energy of the extraction column (Q_E) and regeneration column (Q_R) of the five

widely employed solvents and three candidate green solvents are summarized in Table 6. The detailed operation condition of the eight solvents is tabulated in Table S5.

Table 6. The reboiler heat duties of the extraction column (Q_E) and regeneration column (Q_R) based on five widely used extractive solvents and three candidate green solvents.

Names	Structure	Q_E (kW)	Q_R (kW)	$Q_E + Q_R$ (kW)	Rat oral LD ₅₀ (mg/kg)	Bioconcentration factor (L/kg)
Furfural		919.22	1084.73	2003.96	129	28500
Sulfolane		339.56	2217.88	2557.45	3202	51000
DMSO		1309.15	1178.30	2487.45	1820	48900
DMF		793.33	1804.74	2598.07	2964	93400
NMP		1360.86	1289.77	2650.63	4254	82100
4-methyl furfural		1374.95	1319.23	2694.18	2404	23500
5-methyl furfural		1365.32	1276.51	2641.82	2405	25300
2-hexanone		1490.44	1953.50	3443.94	2490	35300

The total stages of the extractive and regeneration columns are 50 and 40, respectively. The higher the rat oral value of a solvent indicates a higher toxicity. The higher bioconcentration factor of a solvent indicates a greater harm to the ecology.

Additionally, the rat oral and bioconcentration factor information is tabulated in Table 6. The results indicate that there is a trade-off between energy consumption and sustainable performance (such as EH&S properties), where a decrease in energy consumption usually comes at the expense of sustainability. The toxicity of 4-methyl furfural and 5-methyl furfural are reduced by about 95% compared with furfural. The bioconcentration factor of 2-hexanone is reduced by about 62% compared with DMF. The policies worldwide are moving the application of chemical separation processes in the direction of green chemistry⁶. It is worth noting that the reboiler temperature of the extraction and regeneration column of 2-hexanone is lower than 150°C. However, the reboiler temperatures of the extraction and regeneration column of 4-methyl

furfural and 5-methyl furfural are both higher than 150°C. It means that the reboiler using 2-hexanone can use low pressure steam while the reboiler using the other two solvents needs to use medium pressure steam.

6.5 Analysis based on knowledge of the chemistry domain

To make a more intuitive observation, the optimization processes of the three candidate green solvents are visualized in Figure 8. Among the three solvents, the 4-methyl furfural and 5-methyl furfural are the derivatives of furfural. Interestingly, the branching of methyl to the furan ring could significantly reduce the toxicity of furfural. This could be due to the steric effect resulting from the aromatic ring substitution. The dosage of the rat oral of 4-methyl furfural, 5-methyl furfural, and furfural are 2404, 2405, and 129 mg/kg (the higher the better), respectively. The 2-hexanone is obtained by optimizing the structure of DMF. The dialkylation of the carbonyl carbon in DMF can not only improve the selectivity and solution capacity but also reduce the ecological hazards. This is because amide in DMF plays a very pivotal role in the growth and metabolism of microorganisms and can ensure that microbes get enough protein and other important metabolites, thus promoting their growth and reproduction, which could have a negative in the ecology.

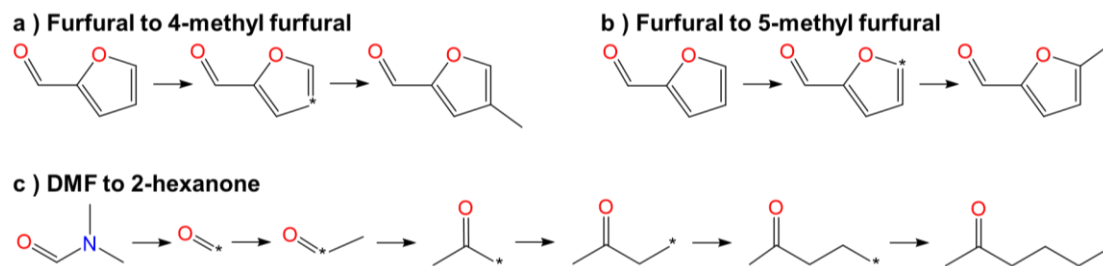


Figure 8. Visualization of the optimization processes of (a) furfural to 4-methyl furfural, (b) furfural to 5-methyl furfural, and (c) DMF to 2-hexanone. * in this figure represents the new fragment attachment (NFA) sites.

In summary, the 4-methyl furfural, 5-methyl furfural, and 2-hexanone can be used as the candidate green solvents to isolate the mixtures of cyclohexane and benzene with extractive distillation. In this contribution, to evaluate the validity of the green solvent multi-objective optimization framework, only 20 molecules are generated from every widely used solvent. More candidate green solvents will be identified if more molecules are optimized and generated for every widely used solvent.

6.6 Molecular fragments analysis

To further explore the relationship between molecular fragments and optimization processes, the fragments are first extracted from the training molecule pairs prepared in Table S4 in the Supplementary Material. The IDACs of these fragments in benzene and cyclohexane are predicted by the proposed IDAC partition models. The selectivity and solution capacity of these fragments are calculated based on the predicted IDACs of these fragments. The detailed information on these fragments is tabulated in Table S7 in the Supplementary Material. The results of the selectivity and solution capacity of these fragments are visualized in Figure 9. In this Figure, molecular fragments with selectivity greater than 3 and solution capacity greater than 0.6 are marked in red. To more intuitively explore the common characteristics between molecular fragments, the molecular structures of the fragments marked in red are visualized in Figure 9. As we can see, most of these visualized molecular fragments are heteroatom-containing aromatic compounds. From the optimization results shown in Figure 6, we can also find that many optimized molecules are modified with these molecular fragments. However, these fragments can easily lead to toxicity and ecological hazards. Therefore, there appears to be a tradeoff between the

separation performance (such as selectivity and solution capacity) and sustainable performance (such as EH&S properties) of the solvents. In this contribution, the proposed green solvent design framework can efficiently balance the trade-off between the separation performance and sustainable performance of the solvents and find the green solvents with multi-constraints.

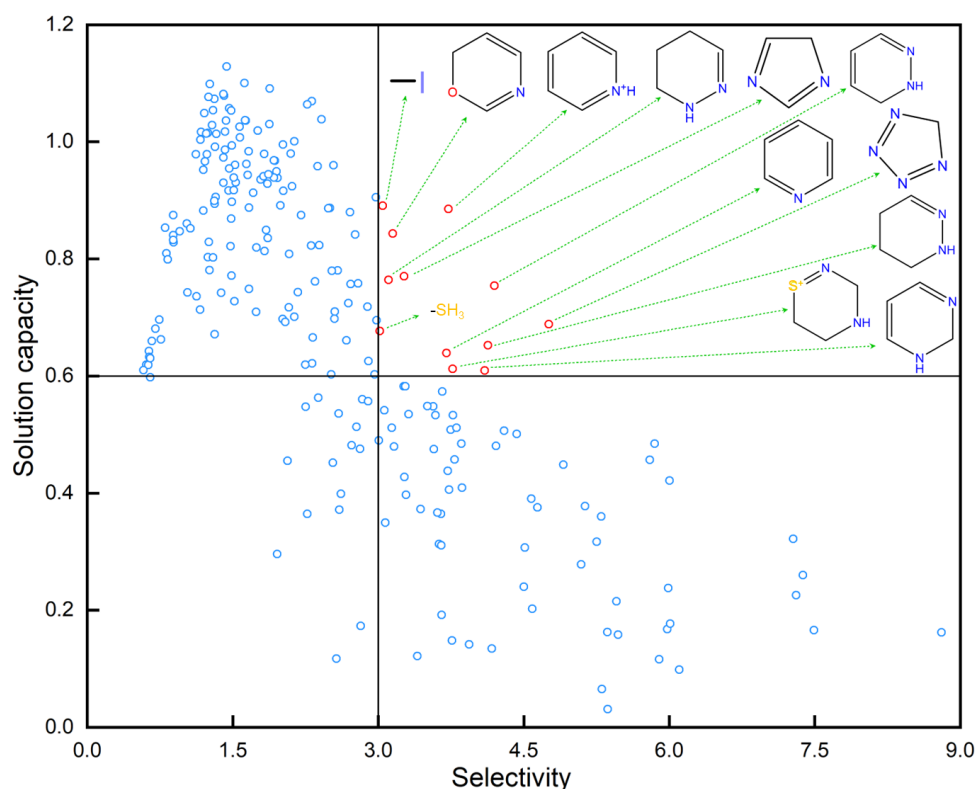


Figure 9. Visualization of the molecular fragment information of selectivity and solution capacity.

7. Conclusions

In this contribution, we propose a molecular multi-objective and multi-scale optimization framework for the design of green solvents fit for extractive distillation that can simultaneously optimize multiple trade-off properties such as selectivity and solution capacity, both related to molecular and process constraints. The molecular multi-objective optimization model relies upon its ability to optimize process

properties rather than molecular properties, as in common computer-aided molecular design approaches. The process properties are short-cut properties of the extractive distillation process, namely selectivity and solution capacity, evaluated *via* infinite dilution activity coefficients (IDAC).

A deep hierarchical molecular multi-objective optimization model is developed to learn the optimization path from our pre-set molecule pairs (M_x , M_y) and generate new solvents by fragment addition or removal. Every two paired molecules in the pre-set molecule pairs are similar in their molecule structures, but the scores of both selectivity and solution capacity of M_y are at least 20% larger than those of M_x . To prepare the molecule pairs, an improved deep learning-based IDAC direct prediction model trained over a COSMO-SAC database is developed for calculating the selectivity and solution capacity of the molecule pairs. The IDAC direct predictive model with the ability to discriminate stereoisomers, achieves a better prediction performance than the IDAC indirect predictive model. As a result, 35,496 molecule pairs are identified that can be used as the training data to train the deep hierarchical molecular multi-objective optimization model. Finally, the proposed IDAC prediction model and molecular multi-objective optimization model are integrated into a green solvent multi-objective and multi-scale optimization framework with EH&S properties and process constraints.

The proposed green solvent multi-objective and multi-scale optimization framework is applied to an extractive distillation process to separate the mixtures of cyclohexane and benzene. The results show that the 4-methyl furfural, 5-methyl furfural, and 2-hexanone can be utilized as candidate green solvents. Among the three

solvents, the 4-methyl furfural and 5-methyl furfural are the derivatives of furfural. Interestingly, the branching of methyl to the furan ring could significantly reduce the toxicity of furfural. This could be due to the steric effect resulting from the aromatic ring substitution. The 2-hexanone is obtained by optimizing the structure of DMF. The dialkylation of the carbonyl carbon in DMF can not only improve the selectivity and solution capacity but also reduce the ecological hazards. This is because amide compounds play a very important role in the growth and metabolism of microorganisms and help microbes get enough protein and other important metabolites, thus promoting their growth and reproduction, which could have a negative impact on the ecology.

Author contributions

Jun Zhang: Conceptualization (lead); data curation (lead); formal analysis (lead); methodology (lead); software (lead); validation (lead); writing – original draft (lead); writing – review and editing (equal). **Qin Wang:** Conceptualization (equal); funding acquisition (equal); methodology (equal); project administration (equal); supervision (equal); writing – review and editing (equal). **Huaqiang Wen:** Formal analysis (equal); methodology (equal); software (equal); validation (equal). **Vincent Gerbaud:** Conceptualization (equal); methodology (equal); writing –review and editing (equal). **Saimeng Jin:** Conceptualization (equal); methodology (equal); writing –review and editing (equal). **Weifeng Shen:** Conceptualization (equal); funding acquisition (lead); methodology (equal); project administration (lead); supervision (lead); writing – original draft (equal); writing – review and editing (lead).

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgements

We acknowledge the financial support provided by the National Natural Science Foundation for Excellent Young Scientists of China (No. 22122802); the National Natural Science Foundation of China (No. 22278044); the Chongqing Science Foundation for Distinguished Young Scholars (No.CSTB2022NSCQ-JQX0021); the Chongqing Innovation Support Key Program for Returned Overseas Chinese Scholars (NO. cx2023002); the Research Foundation of Chongqing University of Science and Technology (No. ckrc2019006).

Date availability statement

The data that supports the findings of this study are available in the supplementary material of this article on <https://zenodo.org/records/10097726>.

References

- 1 J. C. Fromer and C. W. Coley, Computer-Aided Multi-Objective Optimization in Small Molecule Discovery, *Patterns.*, 2023, **4**, 100678.
- 2 X. C. Ma, Q. Zhang, C. He, Q. L. Chen and B. J. Zhang, Computer-aided naphtha liquid–liquid extraction: Molecular reconstruction, sustainable solvent design and multiscale process optimization, *Fuel*, 2023, **334**, 126651.
- 3 S. Chai, Z. Song, T. Zhou, L. Zhang and Z. Qi, Computer-aided molecular design of solvents for chemical separation processes, *Curr. Opin. Chem. Eng.*, 2022, **35**, 100732.
- 4 A. Doolin, R. G. Charles, C. D. Castro, R. G. Rodriguez and M. L. Davies, Sustainable Solvent Selection for the Manufacture of Methylammonium Lead Triiodide (MAPbI₃) Perovskite Solar Cells, *Green Chem.*, 2021, **23**, 2471-2486.
- 5 J. H. Clark, Green chemistry: challenges and opportunities, *Green Chem.*, 1999, **1**, 1-8.
- 6 J. H. Clark, Green Chemistry: Today (and Tomorrow), *Green Chemistry.*, 2006, **8**, 17-21.
- 7 J. Y. Ten, Z. H. Liew, X. Y. Oh, M. H. Hassim and N. Chemmangattuvalappil, Computer-Aided Molecular Design of Optimal Sustainable Solvent for Liquid-Liquid Extraction, *Proc. Integr. Optim.*, 2021, **5**, 269–284.
- 8 Y. S. Lee, A. Galindo, G. Jackson and C. S. Adjiman, Enabling the direct solution of challenging computer-aided molecular and process design problems: Chemical

- absorption of carbon dioxide, *Comput. Chem. Eng.*, 2023, **174**, 108204.
- 9 I. Rodriguez-Donis, S. Thiebaud-Roux, S. Lavoine and V. Gerbaud, Computer-aided product design of alternative solvents based on phase equilibrium synergism in mixtures, *CR CHIM*, 2018, **21**, 606-621.
 - 10 M. Korichi, V. Gerbaud, P. Floquet, A. H. Meniai, S. Nacef and X. Joulia, Computer aided aroma design I-Molecular knowledge framework, *Chem. Eng. Process.*, 2008, **47**, 1902-1911.
 - 11 H. Sun, A Universal Molecular Descriptor System for Prediction of LogP, LogS, LogBB, and Absorption, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 748-757.
 - 12 A. Fredenslund, R. L. Jones and J. M. Prausnitz, Group-contribution estimation of activity coefficients in nonideal liquid mixtures, *AIChE J.*, 1975, **21**, 1086-1099.
 - 13 T. J. Sheldon, M. Folić and C. S. Adjiman, Solvent Design Using a Quantum Mechanical Continuum Solvation Model, *Ind. Eng. Chem. Res.*, 2006, **45**, 1128-1140.
 - 14 J. G. Rittig, K. B. Hicham, A. M. Schweidtmann, M. Dahmen and A. Mitsos, Graph neural networks for temperature-dependent activity coefficient prediction of solutes in ionic liquids, *Comput. Chem. Eng.*, 2023, **171**, 108153.
 - 15 Z. Wang, Y. Su, S. Jin, X. Zhang and J. H. Clark, A novel unambiguous strategy of molecular feature extraction in machine learning assisted predictive models for environmental properties, *Green Chem.*, 2020, **22**, 3867-3876.
 - 16 Z. Wang, Y. Su, W. Shen, S. Jin, J. H. Clark, J. Ren and X. Zhang, Predictive deep learning models for environmental properties: the direct calculation of octanol-water partition coefficients from molecular graphs, *Green Chem.*, 2019, **21**, 4555-4565.
 - 17 T. Zhou, K. McBride, S. Linke, Z. Song and K. Sundmacher, Computer-aided solvent selection and design for efficient chemical processes, *Comput. Chem. Eng.*, 2020, **27**, 35-44.
 - 18 R. Gani, Group contribution-based property estimation methods: advances and perspectives, *Curr. Opin. Chem. Eng.*, 2019, **23**, 184-196.
 - 19 F. Eckert and A. Klamt, Fast solvent screening via quantum chemistry: COSMO-RS approach, *AIChE J.*, 2002, **48**, 369-385.
 - 20 A. Klamt and F. Eckert, COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids, *Fluid Phase Equilibria.*, 2000, **172**, 43-72.
 - 21 S.-T. Lin, *Quantum mechanical approaches to the prediction of phase equilibria: solvation thermodynamics and group contribution methods*, University of Delaware, 2001.
 - 22 S.-T. Lin and S. I. Sandler, A Priori Phase Equilibrium Prediction from a Segment Contribution Solvation Model, *Ind. Eng. Chem. Res.*, 2002, **41**, 899-913.
 - 23 I. H. Bel, E. Mickoleit, C.-M. Hsieh, S.-T. Lin, J. Vrabec, C. Breitkopf and A. Jäger, A Benchmark Open-Source Implementation of COSMO-SAC, *J. Chem. Theory Comput.*, 2020, **16**, 2635-2646.
 - 24 Q. Liu, L. Zhang, K. Tang, L. Liu, J. Du, Q. Meng and R. Gani, Machine learning-based atom contribution method for the prediction of surface charge

- density profiles and solvent design, *AIChE Journal.*, 2021, **67**, e17110.
- 25 T. Mu, J. Rarey and J. Gmehling, Group contribution prediction of surface charge density distribution of molecules for COSMO-SAC, *AIChE J.*, 2009, **55**, 3298-3300.
 - 26 E. Mullins, R. Oldland, Y. A. Liu, S. Wang, S. I. Sandle, C.-C. Chen, M. Zwolak and K. C. Seavey, Sigma-Profile Database for Using COSMO-Based Thermodynamic Methods, *Industrial & Engineering Chemistry Research.*, 2006, **45**, 4389-4415.
 - 27 Y. Su, Z. Wang, S. Jin, W. Shen, J. Ren and M. R. Eden, An architecture of deep learning in QSPR modeling for the prediction of critical properties using molecular signatures, *AIChE Journal.*, 2019, **65**, e16678.
 - 28 J. Zhang, Q. Wang, Y. Su, S. Jin, J. Ren, M. Eden and W. Shen, An accurate and interpretable deep learning model for environmental properties prediction using hybrid molecular representations, *AIChE J.*, 2022, **68**, e17634.
 - 29 F. Jirasek, R. A. Alves, J. Damay, R. A. Vandermeulen, R. Bamler, M. Bortz, S. Mandt, M. Kloft and H. Hasse, Machine learning in thermodynamics: Prediction of activity coefficients by matrix completion, *J. Phys. Chem. Lett.*, 2020, **11**, 981-985.
 - 30 G. Chen, Z. Song, Z. Qi and K. Sundmacher, Neural recommender system for the activity coefficient prediction and UNIFAC model extension of ionic liquid- solute systems, *AIChE J.*, 2021, **67**, e17171.
 - 31 G. Chen, Z. Song and Z. Qi, Transformer-convolutional neural network for surface charge density profile prediction: Enabling high-throughput solvent screening with COSMO-SAC, *Chem. Eng. Sci.*, 2021, **246**, 117002.
 - 32 J. Zhang, Q. Wang and W. Shen, Message-Passing Neural Network Based Multi-Task Deep-Learning Framework for COSMO-SAC based σ -profile and VCOSMO Prediction, *Chem. Eng. Sci.*, 2022, **254**, 117624.
 - 33 R. Gani and E. Brignole, Molecular design of solvents for liquid extraction based on UNIFAC, *Fluid Phase Equilibria.*, 1983, **13**, 331-340.
 - 34 T. Zhou, Z. Song, X. Zhang, R. Gani and K. Sundmacher, Optimal Solvent Design for Extractive Distillation Processes: A Multiobjective Optimization-Based Hierarchical Framework, *Industrial & Engineering Chemistry Research.*, 2019, **58**, 5777-5786.
 - 35 L. Zhang, J. Pang, Y. Zhuang, L. Liu, J. Du and Z. Yuan, Integrated Solvent-Process Design Methodology based on COSMO-SAC and Quantum Mechanics for TMQ (2,2,4-trimethyl-1,2-H-dihydroquinoline) Production, *Chem. Eng. Sci.*, 2020, **226**, 115894.
 - 36 S. Chai, E. Li, L. Zhang, J. Du and Q. Meng, Crystallization solvent design based on a new quantitative prediction model of crystal morphology, *AIChE Journal.*, 2021, **e17499**.
 - 37 J. Heintz, J. P. Belaud, N. Pandya, M. T. D. Santos and V. Gerbaud, Computer aided product design tool for sustainable product development, *Comput. Chem. Eng.*, 2014, **71**, 362-376.
 - 38 A. S. Alshehri, R. Gani and F. You, Deep learning and knowledge-based methods

- for computer-aided molecular design—toward a unified approach: State-of-the-art and future directions, *Comput. Chem. Eng.*, 2020, **141**, 107005.
- 39 A. Graves, Generating sequences with recurrent neural networks, *arXiv preprint arXiv:1308.0850*, 2013.
- 40 V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra and M. Riedmiller, Playing atari with deep reinforcement learning, *arXiv preprint arXiv:1312.5602*, 2013.
- 41 D. P. Kingma and M. Welling, Auto-encoding variational bayes, *arXiv preprint arXiv:1312.6114*, 2013.
- 42 B. Sanchez-Lengeling and A. Aspuru-Guzik, Inverse molecular design using machine learning: Generative models for matter engineering, *Science* 2018, **361**, 360–365.
- 43 A. S. Alshehri and F. You, Deep Learning to Catalyze Inverse Molecular Design, *Chemical Engineering Journal.*, 2022, **444**, 136669.
- 44 W. Jin, R. Barzilay and T. Jaakkola, Junction Tree Variational Autoencoder for Molecular Graph Generation, *arXiv*, 2018, **1802.04364**, arXiv.org e-Print archive.
- 45 W. Jin, R. Barzilay and T. Jaakkola, Hierarchical Generation of Molecular Graphs using Structural Motifs, *arXiv preprint arXiv: 2002.03230v2*, 2020.
- 46 Z. Chen, M. R. Min, S. Parthasarathy and X. Ning, A deep generative model for molecule optimization via one fragment modification, *Nat. Mach. Intell.*, 2021, **3**, 1040-1049.
- 47 J. Wang, C.-Y. Hsieh, M. Wang, X. Wang, Z. Wu, D. Jiang, B. Liao, X. Zhang, B. Yang and Q. He, Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning, *Nat. Mach. Intell.*, 2021, **3**, 914-922.
- 48 D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, S. Golovanov and A. Zhavoronkov, Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models, *Front. Pharmacol.*, 2020, **11**.
- 49 J. Scheffczyk, P. Schäfer, L. Fleitmann, J. Thien, C. Redepenning, K. Leonhard, W. Marquardt and A. Bardow, COSMO-CAMPD: a framework for integrated design of molecules and processes based on COSMO-RS, *Mol. Syst. Des. Eng.*, 2018, **3**, 645-657.
- 50 L. Polte, L. Raßpe-Lange, F. Latz, A. Jupke and K. Leonhard, COSMO-CAMPED – Solvent Design for an Extraction Distillation Considering Molecular, Process, Equipment, and Economic Optimization, *Chem. Ing. Tech.*, 2023, **95**, 416-426.
- 51 S. Kossack, K. Kraemer, R. Gani and W. Marquardt, A systematic synthesis framework for extractive distillation processes, *Chem. Eng. Res. Des.*, 2008, **86**, 781-792.
- 52 V. Gerbaud, I. Rodriguez-Donis, L. Hegely, P. Lang, F. Denes and X. Q. You, Review of extractive distillation. Process design, operation, optimization and control, *Chem. Eng. Res. Des.*, 2019, **141**, 229-271.
- 53 R. Fingerhut, W.-L. Chen, A. Schedemann, W. Cordes, J. r. Rarey, C.-M. Hsieh, J. Vrabec and S.-T. Lin, Comprehensive Assessment of COSMO-SAC Models for

- Predictions of Fluid-Phase Equilibria, *Ind. Eng. Chem. Res.*, 2017, **56**, 9868–9884.
- 54 L. Li, Z. Wen and Z. Wang, *Outlier Detection and Correction During the Process of Groundwater Level Monitoring Base on Pauta Criterion with Self-learning and Smooth Processing. In: Theory, Methodology, Tools and Applications for Modeling and Simulation of Complex Systems*, Springer, Singapore, 2016.
- 55 J. Zhang, Q. Wang, M. Eden and W. Shen, A Deep Learning-based Framework Towards inverse Green Solvent Design for Extractive Distillation with Multi-index Constraints, *Comput. Chem. Eng.*, 2023, **177**, 108335.
- 56 J. Zhang, Q. Wang and W. Shen, Hyper-parameter optimization of multiple machine learning algorithms for molecular property prediction using hyperopt library, *Chin. J. Chem. Eng.*, 2022, **52**, 115-125.
- 57 D. S. Karlov, S. Sosnin, I. V. Tetko and M. V. Fedorov, Chemical space exploration guided by deep neural networks, *RSC Advances.*, 2019, 5151-5157.
- 58 A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis and E. Cibrián-Uhalte, The ChEMBL database in 2017, *Nucleic Acids Res.*, 2017, **45**, D945-D954.
- 59 M. Olivecrona, T. Blaschke, O. Engkvist and H. Chen, Molecular de-novo design through deep reinforcement learning, *J. Cheminformatics*, 2017, **9**, 1-14.
- 60 Z. Abu-Aisheh, R. Raveaux, J. Y. Ramel and P. Martineau, Setubal, 2015.
- 61 L. Sun, Q. Wang, L. Li, J. Zhai and Y. Liu, Design and Control of Extractive Dividing Wall Column for Separating Benzene/Cyclohexane Mixtures, *Ind. Eng. Chem. Res.*, 2014, **53**, 8120–8131.
- 62 Q. Wang, J. Y. Chen, M. Pan, C. He, C. C. He, B. J. Zhang and Q. L. Chen, A new sulfolane aromatic extractive distillation process and optimization for better energy utilization, *Chem. Eng. Proce.*, 2018, **128**, 80-95.
- 63 L. Li, Y. Tu, L. Sun, Y. Hou, M. Zhu, L. Guo, Q. Li and Y. Tian, Enhanced efficient extractive distillation by combining heat-integrated technology and intermediate heating, *Ind. Eng. Chem. Res.*, 2016, **55**, 8837-8847.
- 64 F. M. Lee, Use of organic sulfones as the extractive distillation solvent for aromatics recovery, *Ind. Eng. Chem. Rroce. Des. Dev.*, 1986, **25**, 949-957.
- 65 M. K. Praharaj, A. Satapathy, P. Mishra and S. Mishra, Ultrasonic analysis of intermolecular interaction in the mixtures of benzene with N, N-dimethylformamide and cyclohexane at different temperatures, *J. Chem. Pharm. Res.*, 2013, **5**, 49-56.
- 66 C. Yang, Z. Liu, H. Lai and P. Ma, Thermodynamic properties of binary mixtures of N-methyl-2-pyrrolidinone with cyclohexane, benzene, toluene at (303.15 to 353.15) K and atmospheric pressure, *J. Chem.l Thermodyn.*, 2007, **39**, 28-38.
- 67 Syntelly: Better than chemists can do., <https://syntelly.com>, (accessed 11 Sep., 2023).
- 68 W. Shen, L. Dong, S. Wei, J. Li, H. Benyounes, X. You and V. Gerbaud, Systematic Design of an Extractive Distillation for Maximum-Boiling Azeotropes with Heavy Entrainers, *AIChE Journal.*, 2015, **61**, 3898-3910.
- 69 J. Gu, X. You, C. Tao, L. Jun and G. Vincent, Energy-Saving Reduced-Pressure Extractive Distillation with Heat Integration for Separating the Biazeotropic

Ternary Mixture Tetrahydrofuran-Methanol-Water, *Ind. Eng. Chem. Res.*, 2018, **57**, 13498–13510.

- 70 A. Yang, W. Shen, S. a. Wei, L. Dong, J. Li and V. Gerbaud, Design and control of pressure-swing distillation for separating ternary systems with three binary minimum azeotropes, *AIChE J.*, 2019, **65**, 1281-1293.