Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Chemical Engineering

Review

Molecular machine learning in chemical process design

Jan G Rittig[1,2], Manuel Dahmen[3], Martin Grohe[4], Philippe Schwaller[2,5] and Alexander Mitsos[1,3,6]

Molecular machine learning (ML) has recently demonstrated great potential in (i) predicting properties of pure components and their mixtures, and (ii) exploring the chemical space. We review state-of-the-art molecular ML models, such as graph neural networks and transformers, and discuss research directions for further advancements in chemical process engineering. This includes leveraging molecular ML at the process scale, for example, in process design and optimization formulations, which promises to accelerate the identification of novel molecules and processes. To this end, it will be essential to create design benchmarks and practically validate proposed candidates, possibly in collaboration with the chemical industry.

Addresses
[1] Process Systems Engineering (AVT.SVT), RWTH Aachen University, Aachen, Germany
[2] Laboratory of Artificial Chemical Intelligence (LIAC), Institute of Chemical Sciences and Engineering, EPFL, Lausanne, Switzerland
[3] Forschungszentrum Jülich GmbH, Institute of Climate and Energy Systems ICE-1: Energy Systems Engineering, Jülich, Germany
[4] Lehrstuhl Informatik 7, RWTH Aachen University, Aachen, Germany
[5] National Centre of Competence in Research (NCCR) Catalysis, EPFL, Lausanne, Switzerland
[6] JARA Center for Simulation and Data Science (CSD), Aachen, Germany

Corresponding author: Mitsos, Alexander (amitsos@alum.mit.edu)

Current Opinion in Chemical Engineering 2026, **52**:101239

This review comes from a themed issue on **Artificial intelligence and chemical engineering**

Edited by **Venkat Venkatasubramanian** and **Connor Coley**

Available online xxxx

https://doi.org/10.1016/j.coche.2026.101239

Introduction

Machine learning (ML) has advanced molecular property prediction and design. Over the last years, a variety of ML methods, such as graph neural networks (GNNs) [1], transformers [2], and matrix completion methods (MCMs) [3], have been extensively applied and further developed for predicting properties of molecules and their mixtures. These ML methods have achieved high prediction accuracies, outperforming well-established methods in the field of chemical engineering (ChemE) such as the group contribution method UNIFAC [4] and the quantum mechanics- and statistical thermodynamics-based model COSMO-RS [5] for different properties. Coupling ML with physicochemical knowledge can further greatly enhance or even ensure thermodynamic consistency of the predictions and decrease data required for training [6,7]. Moreover, generative ML models have emerged for computer-aided molecular design (CAMD) [8–10], providing new possibilities for molecular exploration and optimization. Recent studies incorporate experimental validation of ML-designed molecules and target the development of automated experimental molecular design guided by ML, for example, in [11]. Overall, *molecular ML for the transformation toward accelerated molecular design* shows great promise [10].

Within ChemE, it is advantageous and desirable to integrate molecule and process design, cf. [12–14]. Specifically, finding optimal chemical species for a process, that is, molecules such as working fluids, solvents, and products, should be considered as an integrated part of process design. To achieve this, the molecular properties that are relevant for the process are considered as part of the design formulation. Both the molecular and the process structure are considered as degrees of freedom in the design. To date, molecular properties in process models are typically calculated with established thermodynamic property models, for example, NRTL [15] and PC-SAFT [16]. These established models provide very accurate predictions but are limited to molecules for which experimental data is available. In contrast, ML methods enable predictions for molecules not included in model training, for example, see [2,17]. Predictive group contribution methods like UNIFAC show lower accuracy than modern ML approaches, for example, in predicting activity coefficients of binary mixtures [3], and their applicability is limited to molecules for which model parameters are readily available, whereas ML models trained on large data sets typically provide a wider applicability range [3,18]. Further approaches based on quantum mechanics and statistical thermodynamics, such as COSMO-RS [5], can predict a wide range of molecules and properties but (in some cases) have been outperformed by ML, for example, for activity coefficients [17,18] and

solvation free energies [19,20]. However, *process modeling currently lacks state-of-the-art molecular ML models like GNNs*, as such models have not yet been integrated into process simulation software. The identification of suitable chemical species for processes is therefore typically restricted to screening a list of known molecules with readily available property values or thermodynamic model parameters, not making use of recent developments in molecular design with ML. We anticipate that *the integration of ML for molecular property prediction and design with process design and optimization bears large potential and will advance chemical process engineering*.

In the following, we discuss recent concepts in molecular ML and present a perspective on research directions to advance modeling and design at the molecular and process scale in ChemE.

Machine learning for molecules and mixtures

We first provide an overview of molecular ML methods for predicting properties of molecules and mixtures.

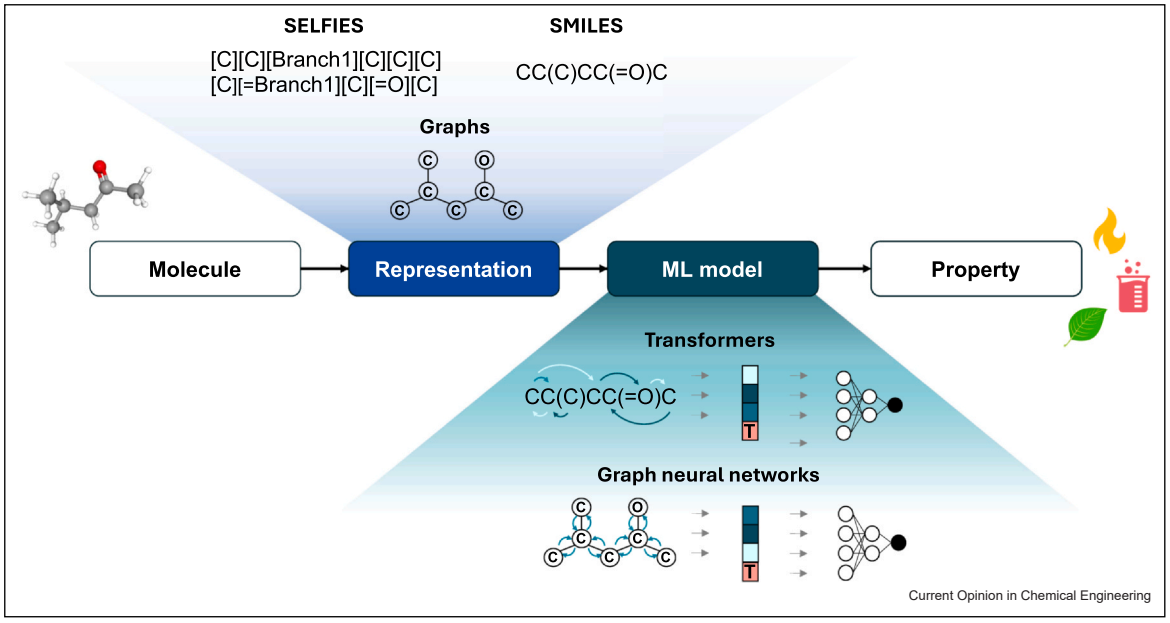
Pure species: from structures to properties

ML models enable to learn physicochemical properties directly from molecular structures of pure species, cf. overviews in [1,23]. We show prominent end-to-end ML approaches for molecular property prediction in Figure 1. The general idea of these end-to-end approaches is first to represent molecules in a machine-readable format. The ML model then encodes this molecular representation into a continuous vector — sometimes referred to as a latent vector or learned molecular fingerprint — in a learnable molecule-to-vector fashion, cf. [24]. The learned vector can be combined with some state information, such as the temperature and pressure, for example, by simple concatenation or a trunk network, cf. [25]. The resulting vector is then mapped to the property of interest, typically by a standard neural network. As the ML model is trained in an end-to-end manner from structure to property [23], the vector representation ideally captures the structural information relevant to the property of interest [26]. The structure-to-property characteristic is the key difference to traditional molecular ML based on molecular descriptors and static fingerprint approaches, like extended connectivity fingerprints [27]. While such traditional molecular ML approaches have a fixed way to transform the structure to a vector representation, the learnable molecule-to-vector encoding in modern ML models is more flexible and allows to achieve state-of-the-art accuracies — if sufficient data is available for training.

In general, two main aspects characterize the type of ML approach: (i) the molecular representation and (ii) the model architecture for the learnable molecular encoding.

(i) *Molecular representation:* Commonly used representations for small molecules are based on

Figure 1



Schematic illustration of molecular ML approaches for property prediction: The molecule is represented in a machine-readable format, for example, strings (here SELFIES [21] and SMILES [22]) or graphs, and then mapped to the property of interest by an ML model, for example, a transformer or a GNN.

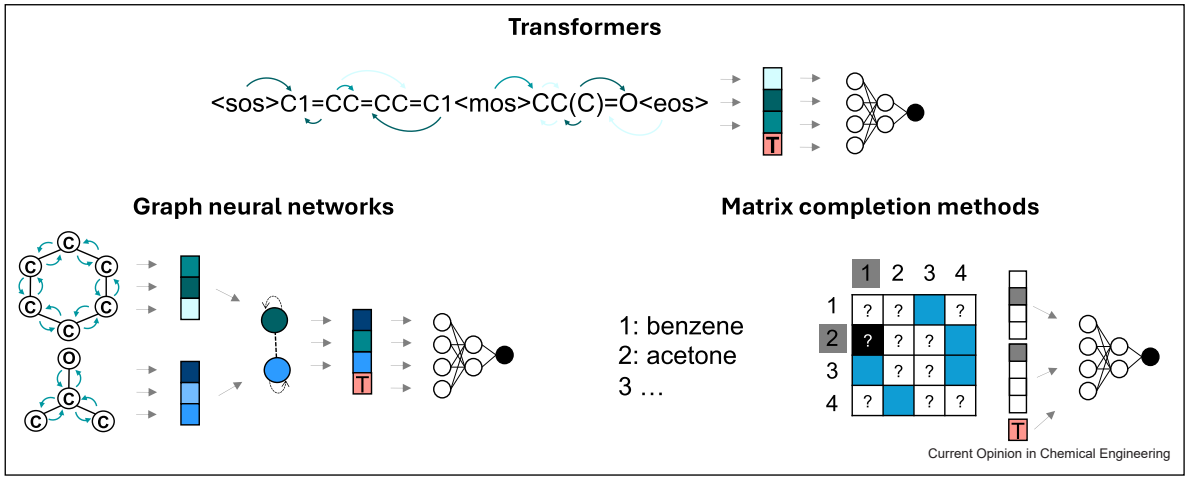
strings, for example, SMILES [22] and SELFIES [21], and on geometry, for example, molecular graphs and point clouds; for detailed overviews, we refer to [21,28–31]. Representing molecules for ChemE applications poses a particular challenge when it comes to larger, more complicated structures, such as polymers and catalysts, as well as multiple interacting structures, as in mixtures and chemical reactions. As such, adaptations of string- and graph-based representations are being actively developed, for example, accounting for the stochastic nature of polymers [32,33].

The used representation determines the level of prior structural information that is provided to the model: Whereas string-based representations and molecular graphs capture the topological structure of molecules and can be enriched by additional descriptors, for example, on stereochemistry, the full spatial information, that is, the arrangement of atoms in 3D space, can be captured in point clouds and geometric graphs, which is highly useful or even necessary for accurate predictions of certain properties, such as electronic ones. However, 3D information is typically not available and thus needs to be calculated with computationally costly quantum mechanical methods. To speed up these calculations, quantum computing and ML interatomic potential approaches are promising and actively researched, cf. overviews in [34–36]. Moreover, representing multiple interacting structures, for example, in mixtures and reactions, in 3D comes with challenges, such as relative positions and orientation, and is thus also a current area of research, see, for example [37]. Herein, we rather focus on string- and graph-based molecular representations, as predictions based on topological structural information have shown high accuracy for many properties relevant to ChemE.

(ii) *ML architecture:* The two most predominant deep-learning architectures for molecular property predictions are transformer architectures [2,38], which originated in natural language processing, and geometric models, mainly GNNs [26,39,40], that respect the natural invariances of graph and spatial representations of the molecules. In the learnable molecular encoding, structural information from the molecular representation is extracted: GNNs extract structural information by passing information along edges in the molecular graph, where edges typically correspond to chemical bonds and the process is referred to as message passing [1,23,40]. That is, GNNs assume that properties are primarily influenced by the set of local atom environments within a molecule. Thus, they come with a strong locality bias, similar to group contribution methods, but with a more flexible, self-evolving character. In contrast, transformers explicitly consider both local and long-range interactions between atoms by the attention mechanism. This also requires inferring chemical principles of bonds and locality from the training data, the chemical 'grammar' [31], typically resulting in the need for pretraining and higher data demands. Notably, from a methodological viewpoint, transformers can be considered as GNNs operating on fully connected graphs and using the attention-based message passing [41]. The two differences in molecular applications between GNNs and transformers lie in (1) the positional encoding typically used in transformers, which can break structural invariances of molecules (e.g. the same molecule can be represented with different SMILES, which lead to different prediction by transformers due to the positional encoding of atoms/tokens), and (2) the inductive bias, that is, whether a molecular property is rather influenced by local atom environments (GNNs) or also by long-range atomic interactions (transformers). Whether transformers are superior to GNNs in exploiting such long-range interactions is actively researched, for example, in [42], and should be further investigated in the molecular context. In this regard, graph transformers that combine the concept of local message passing on graphs with capturing long-range interactions through the attention mechanism are also promising for molecular applications [43–45], but so far less explored in ChemE. As for many molecular properties, the relationship between structure and property is not fully understood; for example, whether local or long-range interatomic effects are relevant, it is advisable to compare the approaches in practical applications.

Molecular ML approaches, such as GNNs and transformers, have recently been extensively applied for the prediction of pure-component properties relevant to ChemE. This includes numerous molecular types, such as small organic molecules [46], polymers [47], and ionic liquids [48], and a variety of properties, such as boiling [23,49] and melting points [50], vapor pressures [49,51,52], density [53], critical micelle concentration [54,55], toxicity [56], and biodegradability [23]. In fact, the prediction capabilities of these ML models have been shown to exceed well-established prediction models based on COSMO, group contributions, and descriptors — in terms of both accuracy and applicability range, see, for example, [49,53], if sufficient data is available for training (typically at least a few hundred data points are needed, cf. [57]). In particular, they enable generalization to novel, unseen components, that is, components for which experimental data is not readily available, given some kind of structural similarity to the molecules used for training. We see the generalization capabilities of ML as most promising for the identification of novel, more sustainable chemical species, see, for example, [11,58,59].

Figure 2



Schematic illustration of molecular ML approaches for predicting mixture properties at the example of a binary mixture of benzene and acetone: transformers considering mixtures as a sequence of SMILES; GNNs taking individual graphs of the molecules within the mixture as input; and MCMs representing a mixture as a combination of one-hot encoded molecular entries along the dimensions, that is, corresponding to the rows and columns in 2D.

Properties of mixtures

Molecular ML models have also been adapted to predict properties of mixtures, as illustrated in Figure 2. Here, next to GNNs and transformers, MCMs have been used. The ways to treat mixtures differ significantly between these methods.

In GNNs, the molecules within a mixture are first encoded to individual vector representations, analogously to pure-component property prediction. The resulting molecular vectors are then aggregated to obtain a vector representing the mixture, the mixture fingerprint, which is then mapped to the mixture property. Several ways of aggregating the molecular fingerprints have been proposed, for example, concatenation or a weighted sum of the molecular vectors, where the weights correspond to the molar fractions [20,60]. Furthermore, mixtures themselves can also be represented by graphs, which allows to capture molecular interactions by applying GNNs before the aggregation step, cf. [18,61,62]; see also [63,64] for further neural interaction/aggregation functions. In comparison to the application of semi-empirical, linear mixing rules based on (predicted) pure component properties, the aggregation of molecule-to-mixture representations within ML models enables learning the mixing effect as a function of the specific molecules within the mixture. Consequently, different mixture behaviors can be predicted.

Transformers treat mixtures as single instances, where the input is typically a sequence of SMILES of the molecules that contains special tokens indicating the start/end of a SMILES, see, for example, Figure 2 and [2]. The model then applies the attention mechanism to the complete sequence, yielding a mixture vector. Notably, transformers do not preserve the order invariance of mixtures, that is, the same mixture can be represented in a different order (e.g. water/ethanol and ethanol/water), but a single sequence implies a fixed order. This issue can be addressed by data augmentation, for example, using multiple sequences for the same mixture with a different order during model training. Here, it would be interesting to investigate architectural adaptions imitating the single molecule encoding and aggregation as in GNNs for mixtures.

MCMs, or more general tensor completion methods, consider mixture property prediction as filling in the missing entries of a matrix, where the dimensions correspond to molecules and the entries to property values. Notably, dimensions can also correspond to states such as temperature. For the completion step, neural networks or Bayesian inference are often used cf. [65,66]. In contrast to GNNs and transformers, MCMs typically do not consider any structural information of the molecules. Rather, the molecules are simply represented as an index in the respective dimension of the matrix/tensor, which is analogous to a one-hot encoding. Thus, the applicability of MCMs is restricted to mixtures that are composed of molecules that occur in the training data set; hence, predicting properties of mixtures with unseen molecules, as with GNNs and transformers, is not possible. To address this issue, it would be interesting to replace the one-hot encoding in MCMs with molecular fingerprints in future work.

All three methods, GNNs, transformers, and MCMs, have been extensively applied to mixture property prediction.

In particular, the activity coefficient of binary mixtures has been targeted, at infinite dilution [2,3,67], varying temperature [18,17,66], varying composition [6,61,62], and all together [7,68,69], resulting in high accuracies of molecular ML, beyond COSMO-RS [5] and UNIFAC [4]. Applications further include properties of mixtures with varying numbers of components, for example, solvation free energies [20], critical micelle concentration [60], and smells [70]. As the combinatorial space of mixtures is vast, especially with an increasing number of components, the high accuracy of molecular ML models is highly promising to accelerate the search for mixtures with desired properties for ChemE applications.

Modeling limitations and practical guidelines

Which end-to-end molecular ML prediction model is most useful depends highly on the property, data availability, and practical requirements for applications. As stated before, GNNs and transformers can be applied to both pure and mixture properties, whereas MCMs are mostly used for mixture properties. For cases where predictions for molecules beyond the training set are needed, hence *generalizing to unseen molecules*, graph- and string-based models, such as GNNs and transformers, are highly useful, whereas MCMs cannot generalize to new molecules.

Furthermore, *computational costs* should be considered when choosing a model. In general, the required computation highly depends on the model structure and hyperparameters. Transformers can be computationally costly in training, as they typically require pretraining on large (simulated) property data sets — here, pretrained models, for example, [71] can be used to save resources. GNNs and MCMs are mostly trained from scratch. The processing of graphs in GNNs can be computationally more costly than using one-hot encodings as in MCMs. Regardless of the model chosen, it is advisable to save resources by using techniques such as early stopping.

Notably, all aforementioned ML models in their pure form — without including any constraints on physical knowledge within the architecture or loss function —, do not *ensure physical consistency* of the predictions; rather, they provide a stochastic estimate based on the data used for training. To ensure physical consistency, the ML models should be enriched by hybrid and physics-informed ML approaches (see next section).

Such physics-enriched approaches can also be helpful in the case of limited *data availability*. Experimental data for properties relevant to ChemE can range from less than a hundred to more than a thousand available data points. Generally, all the mentioned ML models highly depend on the available data for training. In data-scarce cases with less than a few hundred data points), training can be unstable, and predictions might vary a lot between different training runs. In such cases, it is advisable to consider multi-task or transfer learning [19,57] as well as using alternative traditional molecular ML models based on static descriptors and data-driven regression, such as neural networks and XGBoost [72]. End-to-end molecular ML models are likely useful when more data becomes available — in particular, the capabilities of transformers have been shown to scale with the number of parameters and the amount of available data, see, for example [73–75]. In our experience, the quantity and quality of the data are the decisive factors for prediction accuracy rather than the model choice.

In any case, practitioners should always be aware that predictions only provide an initial estimate and that further experimental validation is needed before the molecules of interest can be used in practice. In fact, the extrapolation capabilities of ML models are highly limited; for example, accurately predicting properties of molecules containing atoms not used during training or predicting properties at temperatures beyond the temperature region of the training data points is hardly possible. Generally, the applicability domain and uncertainty quantification of these models should always be taken into account for practical usage.

Research directions

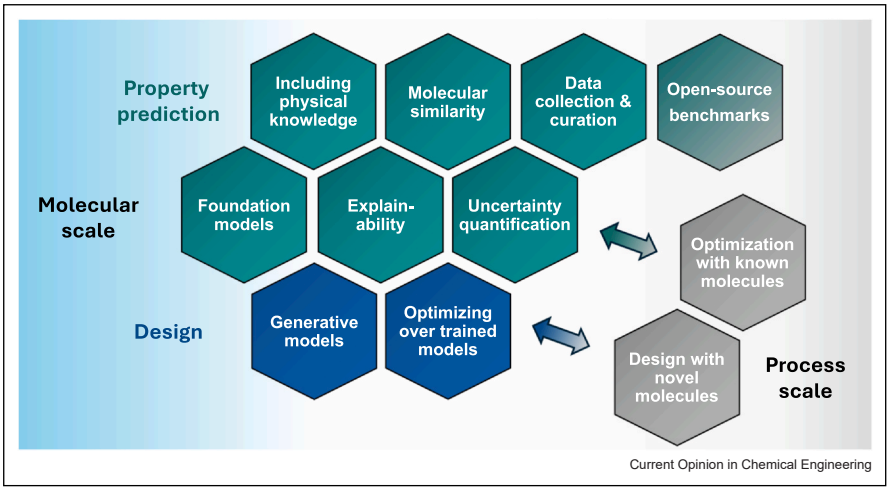
With predictive molecular ML models having shown remarkable results, we see numerous research directions for further advancements and integration with the process scale, as indicated in Figure 3. In the following sections, we identify several research areas that are highly important for the transfer to practical usage and promise further advancements. Furthermore, we argue that molecular ML can substantially contribute to two main objectives of ChemE: the design of more sustainable molecules with desired properties, and the integration into process design and optimization, which we respectively discuss in Sections *Designing molecules with desired properties* and *Toward integration with the process scale*.

Advancing predictive models

We first present research areas that promise advancement in predictive molecular ML models: including physicochemical knowledge, data collection and curation that enable benchmarks, foundation models, explainability, uncertainty quantification, and similarity.

Including physicochemical knowledge will be essential to further advance molecular ML. In recent years, ML models have been adapted to account for physicochemical principles. For example, GNN architectures have been adapted to preserve physical symmetries of molecules, that is, rotational and translational invariance [30,76], account for stereochemical arrangements [77], and consider the influence of the molecular size [78]. Also structural characteristics of certain types of molecules, such as polymers and surfactants, have been used

Figure 3

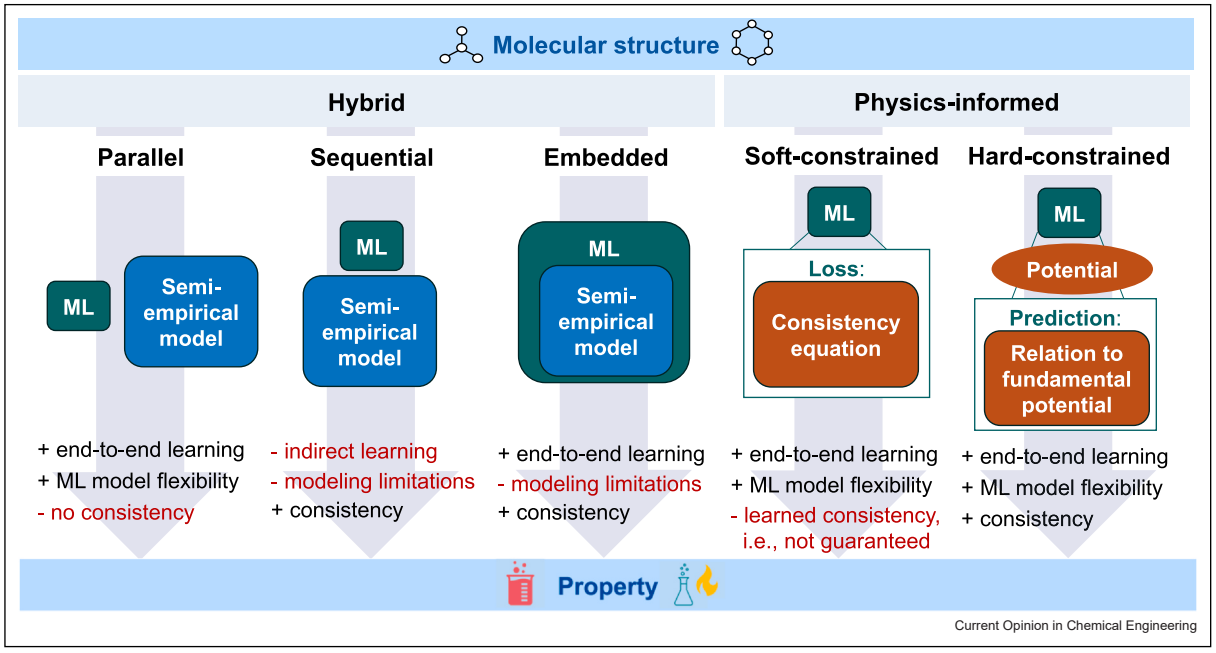


Overview of research areas and directions for molecular ML in chemical process engineering, targeting both the molecular and process scale. Research should advance end-to-end ML models for molecular property prediction and generative ML and optimization approaches for designing molecules with desired properties. Integrating these molecular ML approaches with (ML-based) process modeling, optimization, and design is highly promising.

to refine ML architectures [32,60]. These architectural adaptations can have a significant influence on the consistency and accuracy of the property predictions and should therefore be further explored.

Additionally, hybrid and physics-informed molecular ML approaches have been proposed, as we illustrate in Figure 4, also cf. overview in [79]. *Hybrid models* include the combination of a molecular ML with a semi-

Figure 4



Overview of approaches to combine physicochemical knowledge with molecular ML models for predicting properties based on molecular structures: Hybrid approaches combine molecular ML with semi-empirical models in different forms, that is, in parallel, sequential, or embedded architectures. Physics-informed approaches incorporate physical relations into the training loss for learning consistency (soft-constrained) and use relations to fundamental potentials in the prediction step (hard-constrained).

empirical model in either a sequential, parallel, or embedded way. In the parallel setting, the ML model predicts the error of a semi-empirical model, for example, of COSMO-RS or UNIFAC, as in [67]; the flexibility of ML is preserved but not constrained by any physical knowledge, hence predictions are likely to be physically inconsistent. The sequential approach is characterized by predicting the parameters of a semi-empirical model, while the embedded approach incorporates semi-empirical equations into the ML architecture, such as NRTL or PC-SAFT equations as in [53,68]. Both approaches ensure coherence with the physical knowledge — given that the semi-empirical model is physically consistent; however, the prediction accuracy is constrained by the limitations of the semi-empirical model. We note that the embedded approach should usually be preferred over the sequential approach, as it can be directly trained on property data.

We see two promising research directions building on hybrid approaches: First, applying explainability methods to ML models in the parallel approach can elucidate information on error sources and lead to mechanistic insights. Secondly, the sequential and embedded approach can be used to predict semi-empirical model parameters for molecules for which experimental data is missing and classical parameter fitting is not possible, cf. [53,80,81], so they can directly be utilized in process simulation software, which we explain in more detail in Section *Toward integration with the process scale*.

In contrast to building on semi-empirical models, *physics-informed ML* incorporates algebraic and/or differential relations to fundamental properties. For example, the Helmholtz free energy is a thermodynamic potential from which related properties can be deduced by applying fundamental thermodynamics: intensive properties such as entropy or internal energy are related to first-order derivatives of the Helmholtz free energy; heat capacity and thermal expansion coefficient to second-order derivatives. To incorporate such fundamental relations, two approaches have emerged: soft-constrained and hard-constrained ML. Soft-constrained ML uses relations to fundamental properties as a regularization term in the loss function, that is, the model learns to provide predictions that follow these relations — similar to physics-informed neural networks. Hard-constrained ML describes the concept of embedding relations to fundamental relations into the ML architecture, for example, in [82], or through projection layers, for example, in [83,84], guaranteeing physical consistency. Both approaches have recently been successfully applied for the prediction of thermodynamic properties of fluids and solids [85] and activity coefficients in binary mixtures [6,7,62].

Extending physics-informed molecular ML to further properties and accounting for state transitions [85] will be critical for ensuring physical consistency and achieving higher prediction accuracies while decreasing data demands for training. Physical consistency will also increase acceptance, safety, and trust in molecular ML applications for practitioners.

Property data collection and curation is critical for advancing molecular ML. In our opinion, data scarcity remains the major limiting factor in advancing property prediction for ChemE applications. In fact, sophisticated ML approaches are readily available for molecular applications, but a large fraction of the experimental molecular and mixture property data relevant for ChemE is distributed in numerous literature sources, commercial datasets, and private datasets of chemical companies. Obtaining additional experimental data is naturally costly and labor-intensive, so it is required to increase efforts in leveraging existing data. For publicly available data in the literature, we see large potential of automatic extraction by agentic ML frameworks [86]; yet, human intervention and curation will be required, as reported experimental data can have errors, for example, caused by unit conversion. We advocate for efforts from both academia and industry to assemble such data and make it available for the development of prediction models. For this, the recent collection of data relevant to the chemical science, the ChemPile data set [87], can serve as a blueprint. To additionally utilize proprietary property data, federated learning projects should be initiated that enable and incentivize databank organizations and chemical companies to contribute to the development of ML models without sharing their sensitive data, cf. [88,89]. By scaling the amount of high-quality data that can be utilized for training, the accuracy and applicability domains of ML models will increase.

Benchmarks will catalyze the development of molecular ML models for ChemE applications. Well-defined benchmarks will motivate both the ChemE and the ML community to further develop molecular ML models and reach state-of-the-art accuracies, as can be seen at the prime example of the QM9 dataset [90,91]. In addition, comparing currently available molecular ML models is difficult, as many ChemE-related property data sets used for training and testing are not provided as open-source. We advocate that property prediction benchmarks be created in collaboration with the chemical industry to also include requirements for industrial applications. These benchmarks should cover a wide spectrum of molecule classes and properties relevant for ChemE. Specifically, benchmarks should not only focus on prediction accuracy but also account for physical/thermodynamic consistency. They should further provide multiple test sets to investigate different scenarios, that is, interpolation/extrapolation of state variables, generalization to novel, unseen molecules, etc. A notable recent example is the CheMixHub benchmark for

predicting mixture properties [70], which includes 11 different properties and various test splits. Open-source property prediction benchmarks can thereby direct the development of new molecular ML approaches to meet industrial needs and criteria for practical applications.

Foundation models promise to advance molecular ML by training on large amounts of property data, so that they can generalize and be fine-tuned on specific property prediction tasks, even if only little data is available [92]. To date, the typical approach in molecular ML is to train predictive models from scratch, that is, for a property prediction task at hand, data is collected and then an ML model is trained, often using readily implemented molecular ML frameworks like chemprop [39,93]. This can be challenging as property data is often scarce in ChemE applications. To address this issue, mainly self-supervised, transfer, multi-task, and multi-fidelity learning approaches have been investigated in the molecular domain, for example, in [19,43,71,94]. While in some cases, these methods can lead to improvements in the accuracy and applicability range of the prediction, the experimental data used is rather small and covers only a few properties, which limits generalization. Further cases show that combining property data as in multi-task models can also decrease model performance, as the optimization during training becomes more difficult, see for example, [39,95]. We argue that using additional well-curated data should generally increase — at least not harm — model performance, so further developments in scaling molecular ML architectures and improving training procedures, for example, recently proposed task-specific early stopping [96], are needed to utilize the information and relationships hidden in molecular property data sets.

Recently, a few studies have assembled large molecular data sets and trained transformers and GNNs on molecular property prediction, for example, [44,97,98]. However, these studies mostly focus on properties relevant for biological and pharmaceutical applications, whereas ChemE lacks large data sets. Thus, we believe that assembling large data sets with properties relevant for ChemE by combining data for different properties bears large potential for ML. In fact, we hypothesize that increased diversity of molecular classes and properties could enable ML to exploit and uncover chemical patterns beyond simple property correlations, which will facilitate generalization; whether this will be in the sense of a foundation model remains to be explored. In particular, we see great potential in including *fundamental properties*, such as Gibbs and Helmholtz free energies, as these provide insights into relations between different properties and enable consistent predictions [76]. Overall, combining molecular ML models with fundamental property relations and training them on large-scale property datasets can advance property prediction in small data applications and increase generalization capabilities.

Explainability promises to uncover unknown molecular structure-property relationships. Explaining and interpreting predictions of molecular ML models has been actively researched in recent years, for example, by investigating model sensitivities through gradient-based or counterfactual methods, cf. [99,100]. So far, most approaches focus on explaining local, single-instance predictions, that is, for a given molecule, which is useful in validating whether an ML model correctly identifies structural parts of the molecule that are known to influence a property. We advocate for also focusing on systematically explaining ML predictions on a global model level [101], that is, finding generalizable structure-property relations that hold for a diverse collection of molecules. For example, subgraphs representing molecular motifs can be clustered and analyzed with large language models as in [102], or learned molecular vectors can be utilized in hierarchical clustering to identify property-specific molecular classes [103]. Such research should also distinguish between abductive ("Which structural parts support a certain property prediction?") and contrastive relations ("Which structural parts need to be changed to get a different property prediction?") [104,105].

A significant aspect to consider here is that ML models are highly overparameterized and capture nonlinearities in large data sets that enable reaching accuracies beyond mechanistic models. So, we hypothesize that related phenomena are difficult to translate to high-level explanations for humans. Important questions that should be addressed here are: What are the explanations for the accuracy gains beyond mechanistic models? To what extent can we achieve a mechanistic understanding with ML models, and are explainability and high accuracy conflicting objectives at some point?

Uncertainty quantification is highly important for practical applications of molecular ML. Numerous uncertainty quantification methods have been investigated for molecular ML, including similarity- and ensemble-based methods, mean-variance estimation, and conformal prediction, with ambiguous results regarding superiority and usability, cf. overview in [106]. The field continues to be actively researched, and promising methods are being proposed frequently for ChemE applications, for example, based on architecture search [107] or stochastic gradient Hamiltonian Monte Carlo [108]. A major challenge here is to decompose the uncertainty in the epistemic part caused by the model and the aleatoric part [109], that is, the uncertainty inherent to the property data due to different experimental setups, instruments, reporting, etc. To elucidate model uncertainties, quantification methods have very recently been coupled

with explainability approaches, see, for example, [110,111].

It will be particularly interesting to test developed uncertainty quantification methods on property prediction benchmarks created for ChemE. We also stress that it is promising to test and further develop these methods for multi-task models, and ultimately foundation models, since this could help to infer uncertainty relationships between different properties and facilitate identifying erroneous data points.

Molecular similarity based on molecular fingerprint vectors can reveal novel chemical relations that are learned by molecular ML models. The concept of similarity is frequently used in analyzing molecular latent spaces. Specifically, the learned fingerprint vectors in molecular ML models are reduced to a few (typically two) dimensions and visualized in a human-understandable way. Then, distances between individual vectors can be interpreted as molecular similarity specific to the physicochemical property the molecular ML model is trained on, potentially revealing chemical insights, for example, clusters of molecular classes [19]. Furthermore, similarity-based approaches can be used to assess model uncertainties; that is, predictions are assumed to have higher accuracy for molecules that are encoded into a fingerprint vector close to those of the molecules used for training, cf. [106]. Analyzing molecular similarity is thus highly related to explainability and uncertainty quantification research.

We see further need in investigating molecular similarity in learned fingerprint/latent spaces, as the learnable molecule-to-vector encoding is an essential part of molecular ML models — distinguishing them from established, static fingerprint approaches. The following specific questions should be addressed: What are the differences in the distance of the learned fingerprint vectors depending on the property to be predicted? Is the fingerprint vector space actually interpretable, given its typical high dimensionality, and how does the number of dimensions influence the similarity? How does the learned property-oriented similarity relate to structural similarity based on static fingerprints, as well as to more abstract concepts, such as string or graph similarity? Lastly, it would be interesting to research similarity in multi-task models. For example, do shared model layers capture high-level chemical concepts that are relevant to subsets or all of the considered properties, and how do the fingerprint vectors change in property-specific layers? Addressing these questions would greatly increase the understanding and interpretability of molecular ML models.

Overall, we find many promising research directions that have the potential to increase the predictive capabilities of molecular ML and uncover novel chemical insights. In particular, we see the need for strong collaboration between academia and industry to improve the development and reliability of molecular ML models, which will eventually lead to their practical application on an industrial scale.

Designing molecules with desired properties

While molecular ML models enable property prediction, it is desirable to identify molecules with optimal properties for specific ChemE applications, referred to as CAMD. For this, we identify two actively investigated approaches: (i) generative ML models and (ii) deterministic global optimization over molecular ML models.

Generative ML models propose new molecules with desired properties by learning from existing molecular structures. Notably, they can explore the chemical space beyond established CAMD approaches in ChemE, for example, based on structure enumeration or group contribution models embedded into optimization formulations, which are restricted to a combination of functional groups [112]. Over the last years, numerous generative molecular ML models have been proposed, including variational autoencoders (VAEs), reinforcement learning (RL), generative adversarial networks, and recently diffusion- and flow-based models, cf. overviews in [8,10].

In ChemE, generative models have recently been applied for identifying promising molecules as fuels [113,114], polymers [115], and solvent [116,117]. Importantly, ML-designed molecules need to be synthesizable and chemically stable for usage in practical applications. As such, generative ML models have been extended to also account for synthesizability and constraints on molecular motifs/building blocks [116,118,119], making them particularly promising to accelerate molecular discovery. Indeed, experimental validation remains critical and has been demonstrated in some recent ML-based CAMD works [11,113]. Future works in ChemE should focus on combining CAMD with (automated) experimental testing to create iterative molecular discovery pipelines.

Global optimization with molecular ML models embedded enables finding molecules with globally optimal properties. Specifically, ML models that are trained to predict molecular properties can be embedded into optimization formulations for molecular design. As such, the ML model weights are fixed, and the prediction is optimized as a function of the inputs, that is, the molecular structure is the degree of freedom; notably, the prediction can also be considered as a constraint in a design formulation. For example, trained GNNs have been embedded into molecular design formulations, which enables to find global optimal molecules as

predicted by the GNN using deterministic solvers [120,121]. For this, two major challenges arise: additional constraints need to be formulated to restrict the search space of molecular structures to chemically valid molecular graphs, and the highly nonlinear GNN layers are part of the problem formulation, making solving computationally costly and currently impractical for molecules with more than a few atoms [120–122]. As an alternative approach that circumvents these challenges, VAEs for molecule generation can jointly be trained with neural networks for property prediction on the VAE's latent space, which allows to only consider the neural network in the molecular design formulation, cf. [122,123]. This approach, however, comes with the additional computational costs and difficulty of training a VAE jointly with a neural network.

Overall, optimizing over molecular ML models is highly promising for molecular design, as it enables finding optimal molecules (as predicted). We advocate for further research in this area, including the embedding of other molecular ML models into optimization formulations, such as transformers and MCMs. Embedding molecular ML models into optimization formulations will also be of major importance in process design, which we will discuss next.

Toward integration with the process scale

The integration of molecular ML into the process scale of ChemE is still in its infancy. In fact, molecular ML is rarely used for: (i) predicting the properties of chemical species used in chemical processes in the context of process modeling and optimization; and (ii) designing molecules and mixtures as an integrated part of process design, known as computer-aided molecular and process design (CAMPD), cf. overviews in [124–128]. In fact, CAMPD approaches typically embed molecular fragmentation and group contribution methods in optimization formulations, limiting the molecular and process design space.

We advocate for the integration of molecular ML into process models, for example, for modeling thermodynamic properties and sustainability factors. Here, we distinguish two scenarios, considering (i) known molecules in practical use, and (ii) generalizing to novel molecules.

Known molecules in practical use within chemical processes typically come with readily available property data. This means that established semi-empirical models, for example, based on equation-of-state approaches, or surrogate models, such as polynomials and shallow neural networks, can be fitted to this data. These models typically provide reasonably accurate property predictions for process optimization and design, without the need for advanced molecular ML approaches.

However, in cases where the property data or semi-empirical/surrogate models are limited to specific state ranges, for example, in terms of temperature and pressure, process optimization becomes restricted. Molecular ML models trained on a diverse set of molecules with corresponding properties can provide state-dependent predictions for wider ranges than a model fitted only on the property data of a single molecule (or mixture) of interest. As such, molecular ML models need to be embedded into process model formulations, for example, using tools such as OMLT [129] or MeLOn [130], which will require model adjustments and might cause additional computational costs, cf. for example, [120,121]. Alternatively, semi-empirical model parameters for the molecule (or mixture) of interest can be fitted to predictions of molecular ML models or extracted from hybrid molecular ML architectures that are trained on a more diverse set of molecules and wider state ranges, cf. [53,80] and Section *Machine learning for molecules and mixture*. These parameters can then be directly used in process modeling without any additional model adjustments. Therefore, molecular ML can expand process optimization and design to include wider operating ranges, potentially leading to increased process efficiencies.

Generalizing to novel molecules with desired properties for chemical processes is highly desirable. For this, process performance indicators can be included in molecular design objectives (and constraints). For the example of finding suitable solvents in separation processes, properties such as solubility or partition coefficients can be directly optimized [59,116] or used for predicting process performance with data-driven surrogate models, cf. [131]. Further properties, for example, accounting for the environmental impact of the proposed molecules, can also be included in the design. Predictive and generative molecular ML models can then be employed and accelerate the identification of novel, sustainable solvents, reactants, catalysts, etc. that optimize these process performance indicators, whereas subsequent rigorous process optimization and experimental validation remain critical.

Ultimately, we anticipate that ML-driven CAMPD will play an important role in ChemE, that is, molecules and processes are designed simultaneously through ML. In addition to embedding molecular ML models into process models, similarly to the approach for known molecules described above, the molecular structure becomes a degree of freedom, making optimization much more challenging and thus requiring further research. To circumvent embedding equations of molecular ML models into optimization formulations, sequential ML-CAMPD workflows can be employed, as very recently proposed in [128] for the design of solvent-antisolvent mixtures and crystallization processes. Specifically, molecular ML can

be utilized in an iterative manner by (1) proposing molecular structures by molecular design algorithms, (2) predicting the properties of the proposed structures by predictive ML models, and (3) solving a process design formulation using these predicted properties, which then serves as a feedback for the design algorithm in (1) [128], thereby integrating process design goals into molecular design.

Another particularly promising direction is to couple molecular ML with recently proposed generative ML approaches for process design [132]. Such ML-based process design approaches, mostly based on RL, so far only include process variables as part of the design space [133,134]. Developing generative ML methods that enable the simultaneous design of molecules and processes bears large potential in automating and advancing CAMPD.

Concluding remarks

ML has advanced molecular property prediction and design in ChemE by learning from data on molecules and mixtures. We hypothesize that there are more chemical relationships hidden in these data than current molecular ML models have learned. Data collection and curation is needed so that ML models can leverage and reveal these relationships through advanced model architectures that are based on physicochemical knowledge and model-level explainability. Furthermore, coupling molecular ML with the process scale will accelerate the identification of novel, more sustainable molecules and mixtures that also lead to more efficient processes. It is of major importance that academia closely collaborates with the chemical industry to further advance molecular ML models and establish benchmarks for practical application in process design and optimization.

Future work should focus on integrating generative ML for the molecular and process scale. To this end, we see great potential in multi-agent frameworks [31,135,136] for orchestrating and automating ChemE design tasks.

Data Availability

No data were used for the research described in the article.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 466417970 - within the Priority Programme "SPP 2331: Machine Learning in Chemical Engineering".

This work was also performed as part of the Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE).

The project was also funded by the European Union (ERC, SymSim, 101054974). This work was further funded by the Swiss Confederation under State Secretariat for Education, Research and Innovation SERI, participating in the European Union Horizon Europe project ILIMITED (101192964). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Funding by the Werner Siemens Foundation within the WSS project of the century "catalaix" is acknowledged.

MD and AM received funding from the Helmholtz Association of German Research Centers.

PS acknowledges support from the NCCR Catalysis (grant number 225147), a National Centre of Competence in Research funded by the Swiss National Science Foundation.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Reiser P, Neubert M, Eberhard A, Torresi L, Zhou C, Shao C, Metni H, van Hoesel C, Schopmans H, Sommer T, Friederich P: **Graph neural networks for materials science and chemistry**. *Commun Mater* 2022, **3**:93.

2. Winter B, Winter C, Schilling J, Bardow A: **A smile is all you need: predicting limiting activity coefficients from SMILES with natural language processing**. *Digit Discov* 2022, **1**:859-869.

3. Jirasek F, Alves RAS, Damay J, Vandermeulen RA, Bamler R, Bortz M, Mandt S, Kloft M, Hasse H: **Machine learning in thermodynamics: Prediction of activity coefficients by matrix completion**. *J Phys Chem Lett* 2020, **11**:981-985.

4. Fredenslund A, Jones RL, Prausnitz JM: **Group-contribution estimation of activity coefficients in nonideal liquid mixtures**. *AIChE J* 1975, **21**:1086-1099.

5. Klamt A, Eckert F, Arlt W: **COSMO-RS: an alternative to simulation for calculating thermodynamic properties of liquid mixtures**. *Annu Rev Chem Biomol Eng* 2010, **1**:101-122.

6. • Rittig JG, Mitsos A: **Thermodynamics-consistent graph neural networks**. *Chem Sci* 2024, **15**:18504-18512.
This work demonstrates how fundamental thermodynamics can be integrated with molecular machine learning models such as GNNs, r esulting in highly accurate predictions and ensuring thermodynamic consistency.

7. Specht T, Nagda M, Fellenz S, Mandt S, Hasse H, Jirasek F: **HANNA: hard-constraint neural network for consistent activity coefficient prediction**. *Chem Sci* 2024, **15**:19777-19786.

8. Bilodeau C, Jin W, Jaakkola T, Barzilay R, Jensen KF: **Generative models for molecular discovery: recent advances and challenges**. *Wiley Interdiscip Rev Comput Mol Sci* 2022, **12**:e1608.

9. Elton DC, Boukouvalas Z, Fuge MD, Chung PW: **Deep learning for molecular design - a review of the state of the art**. *Mol Syst Des Eng* 2019, **4**:828-849.

10. Du Y, Jamasb AR, Guo J, Fu T, Harris C, Wang Y, Duan C, Liò P, Schwaller P, Blundell TL: **Machine learning-aided generative molecular design**. *Nat Mach Intell* 2024, **6**:589-604.

11. Koscher BA, Canty RB, McDonald MA, Greenman KP, McGill CJ, Bilodeau CL, Jin W, Wu H, Vermeire FH, Jin B, *et al.*: **Autonomous, multiproperty-driven molecular discovery: From predictions to measurements and back**. *Science* 2023, **382**:eadi1407.

12. Bardow A, Steur K, Gross J: **Continuous-molecular targeting for integrated solvent and process design**. *Ind Eng Chem Res* 2010, **49**:2834-2840.

13. Zhang L, Babi DK, Gani R: **New vistas in chemical product and process design**. *Annu Rev Chem Biomol Eng* 2016, **7**:557-582.

14. Burger J, Papaioannou V, Gopinath S, Jackson G, Galindo A, Adjiman CS: **A hierarchical method to integrated solvent and process design of physical CO_2 absorption using the SAFT-γ Mie approach**. *AIChE J* 2015, **61**:3249-3269.

15. Renon H, Prausnitz JM: **Local compositions in thermodynamic excess functions for liquid mixtures**. *AIChE J* 1968, **14**:135-144.

16. Gross J, Sadowski G: **Perturbed-chain SAFT: an equation of state based on a perturbation theory for chain molecules**. *Ind Eng Chem Res* 2001, **40**:1244-1260.

17. Rittig JG, BenHicham K, Schweidtmann AM, Dahmen M, Mitsos A: **Graph neural networks for temperature-dependent activity coefficient prediction of solutes in ionic liquids**. *Comput Chem Eng* 2023, **171**:108153.

18. •• SanchezMedina EI, Linke S, Stoll M, Sundmacher K: **Gibbs–Helmholtz graph neural network: capturing the temperature dependency of activity coefficients at infinite dilution**. *Digit Discov* 2023, **2**:781-798.
This work collects the largest dataset on temperature-dependent activity coefficients at infinite dilution and trains a highly accurate GNN model that incorporates the Gibbs-Helmholtz equation.

19. Vermeire FH, Green WH: **Transfer learning for solvation free energies: from quantum chemistry to experiments**. *Chem Eng J* 2021, **418**:129307.

20. Leenhouts RJ, Morgan N, Allbrahim E, Green WH, Vermeire FH: **Pooling solvent mixtures for solvation free energy predictions**. *Chem Eng J* 2025, **513**:162232.

21. Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A: **Self-referencing embedded strings (selfies): a 100% robust molecular string representation**. *Mach Learn Sci Technol* 2020, **1**:045024.

22. Weininger D: **SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules**. *J Chem Inf Comput Sci* 1988, **28**:31-36.

23. Rittig JG, Gao Q, Dahmen M, Mitsos A, Schweidtmann AM: **Graph neural networks for the prediction of molecular structure–property relationships**. In *Machine Learning and Hybrid Modelling for Reaction Engineering*. Edited by Zhang D, DelRíoChanona EA. Royal Society of Chemistry; 2023:159-181.

24. Grohe M: **word2vec, node2vec, graph2vec, x2vec: Towards a theory of vector embeddings of structured data.** In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*; 2020: 1–16.

25. Pavšek, J, Mitsos, A, Dahmen, M, Tan, TX, & Rittig, JG: DeepEOSNet: Capturing the dependency on thermodynamic state in property prediction tasks. 2025, arXiv preprint arXiv:2509.17018.

26. Coley CW, Barzilay R, Green WH, Jaakkola TS, Jensen KF: **Convolutional embedding of attributed molecular graphs for physical property prediction**. *J Chem Inf Model* 2017, **57**:1757-1772.

27. Rogers D, Hahn M: **Extended-connectivity fingerprints**. *J Chem Inf Model* 2010, **50**:742-754.

28. Wigh DS, Goodman JM, Lapkin AA: **A review of molecular representation in the age of machine learning**. *Wiley Interdiscip Rev Computationa Mol Sci* 2022, **12**:e1603.

29. Atz K, Grisoni F, Schneider G: **Geometric deep learning on molecular representations**. *Nat Mach Intell* 2021, **3**:1023-1032.

30. A. Duval, S.V. Mathis, C.K. Joshi, V. Schmidt, S. Miret, F.D. Malliaros, T. Cohen, P. Lio, Y. Bengio, M. Bronstein: **A Hitchhiker's Guide to Geometric GNNs for 3D Atomic Systems**. arXiv preprint arXiv:2312.07511. doi:10.48550/arXiv.2312.07511.

31. • Alampara N, Aneesh A, Ríos-García M, Mirza A, Schilling-Wilhelmi M, Aghajani AA, Sun M, Prastalo G, Jablonka KM: **General-purpose models for the chemical sciences: LLMs and beyond**. *Chem Rev* 2026, **126**:2484-2549, https://doi.org/10.1021/acs.chemrev.5c00583.
This work provides a comprehensive review and future research directions on large language models and agentic AI for chemical sciences, including molecular data collection, property prediction, and design.

32. Aldeghi M, Coley CW: **A graph representation of molecular ensembles for polymer property prediction**. *Chem Sci* 2022, **13**:10486-10498.

33. Lin T-S, Coley CW, Mochigase H, Beech HK, Wang W, Wang Z, Woods E, Craig SL, Johnson JA, Kalow JA, *et al.*: **BigSMILES: a structurally-based line notation for describing macromolecules**. *ACS Cent Sci* 2019, **5**:1523-1531.

34. Cao Y, Romero J, Olson JP, Degroote M, Johnson PD, Kieferová M, Kivlichan ID, Menke T, Peropadre B, Sawaya NP, *et al.*: **Quantum chemistry in the age of quantum computing**. *Chem Rev* 2019, **119**:10856-10915.

35. Anstine DM, Zubatyuk R, Isayev O: **AIMNet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs**. *Chem Sci* 2025, **16**:10228-10244.

36. Jacobs R, Morgan D, Attarian S, Meng J, Shen C, Wu Z, Xie CY, Yang JH, Artrith N, Blaiszik B, *et al.*: **A practical guide to machine learning interatomic potentials–status and future**. *Curr Opin Solid State Mater Sci* 2025, **35**:101214.

37. van Gerwen P, Briling KR, Bunne C, Somnath VR, Laplaza R, Krause A, Corminboeuf C: **3DReact: Geometric deep learning for chemical reactions**. *J Chem Inf Model* 2024, **64**:5771-5785.

38. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I: **Attention is all you need**. In *Advances in Neural Information Processing Systems*. Edited by Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R. 30 Curran Associates, Inc.; 2017, https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

39. Heid E, Greenman KP, Chung Y, Li S-C, Graff DE, Vermeire FH, Wu H, Green WH, McGill CJ: **Chemprop: a machine learning package for chemical property prediction**. *J Chem Inf Model* 2024, **64**:9-17.

40. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE: **Neural message passing for quantum chemistry.** In *Proceedings of the 34th International Conference on Machine Learning*. Edited by Precup D, Teh YW. PMLR; 2017, 70:1263-1272. https://proceedings.mlr.press/v70/gilmer17a.html.

41. C.K. Joshi, **Transformers Are Graph Neural Networks.** arXiv preprint arXiv:2506.22084. https://doi.org/10.48550/arXiv.2506.22084.

42. J. Tönshoff, M. Ritzert, E. Rosenbluth, M. Grohe: **Where Did the Gap Go? Reassessing the Long-range Graph Benchmark**, arXiv preprint arXiv:2309.00367. https://doi.org/10.48550/arXiv.2309.00367.

43. Rong Y, Bian Y, Xu T, Xie W, Wei Y, Huang W, Huang J: **Self-supervised graph transformer on large-scale molecular data**. *Adv Neural Inf Process Syst* 2020, **33**:12559-12571.

44. Sypetkowski M, Wenkel F, Poursafaei F, Dickson N, Suri K, Fradkin P, Beaini D: **On the scalability of gnns for molecular graphs**. *Adv Neural Inf Process Syst* 2024, **37**:19870-19906.

45. Anselmi M, Slabaugh G, Crespo-Otero R, DiTommaso D: **Molecular graph transformer: stepping beyond alignn into long-range interactions**. *Digit Discov* 2024, **3**:1048-1057.

46. Dou B, Zhu Z, Merkurjev E, Ke L, Chen L, Jiang J, Zhu Y, Liu J, Zhang B, Wei G-W: **Machine learning methods for small data challenges in molecular science**. *Chem Rev* 2023, **123**:8736-8780.

47. Ge W, De Silva R, Fan Y, Sisson SA, Stenzel MH: **Machine learning in polymer research**. *Adv Mater* 2025, **37**:2413695.

48. Song Z, Chen J, Cheng J, Chen G, Qi Z: **Computer-aided molecular design of ionic liquids as advanced process media: a review from fundamentals to applications**. *Chem Rev* 2023, **124**:248-317.

49. Hoffmann M, Hasse H, Jirasek F: **GRAPPA–a hybrid graph neural network for predicting pure component vapor pressures**. *Chem Eng J Adv* 2025,100750.

50. Sivaraman G, Jackson NE, Sanchez-Lengeling B, Vázquez-Mayagoitia Á, Aspuru-Guzik A, Vishwanath V, Pablo JJDe: **A machine learning workflow for molecular analysis: application to melting points**. *Mach Learn Sci Technol* 2020, **1**:025015.

51. Lansford JL, Jensen KF, Barnes BC: **Physics–informed transfer learning for out–of–sample vapor pressure predictions**. *Propellants Explos Pyrotech* 2023, **48**:e202200265.

52. Santana VV, Rebello CM, Queiroz LP, Ribeiro AM, Shardt N, Nogueira IB: **PUFFIN: a path-unifying feed-forward interfaced network for vapor pressure prediction**. *Chem Eng Sci* 2024, **286**:119623.

53. Winter B, Rehner P, Esper T, Schilling J, Bardow A:
• **Understanding the language of molecules: predicting pure component parameters for the PC-SAFT equation of state from SMILES**. *Digit Discov* 2025, **4**:1142-1157, https://doi.org/10.1039/D4DD00077C.
This work embeds the PC-SAFT equation of state into a transformer architecture, demonstrating end-to-end learning of molecular properties through thermodynamic model.

54. Qin S, Jin T, Van Lehn RC, Zavala VM: **Predicting critical micelle concentrations for surfactants using graph convolutional neural networks**. *J Phys Chem B* 2021, **125**:10610-10620.

55. Brozos C, Rittig JG, Bhattacharya S, Akanny E, Kohlmann C, Mitsos A: **Graph neural networks for surfactant multi-property prediction**. *Colloids Surf A Physicochem Eng Asp* 2024, **694**:134133.

56. Seal S, Mahale M, García-Ortegón M, Joshi CK, Hosseini-Gerami L, Beatson A, Greenig M, Shekhar M, Patra A, Weis C, *et al.*: **Machine learning for toxicity prediction using chemical structures: pillars for success in the real world**. *Chem Res Toxicol* 2025, **38**:759-807.

57. Schweidtmann AM, Rittig JG, König A, Grohe M, Mitsos A, Dahmen M: **Graph neural networks for prediction of fuel ignition quality**. *Energy Fuels* 2020, **34**:11395-11407.

58. Peng J, Schwalbe-Koda D, Akkiraju K, Xie T, Giordano L, Yu Y, Eom CJ, Lunger JR, Zheng DJ, Rao RR, *et al.*: **Human-and machine-centred designs of molecules and materials for sustainability and decarbonization**. *Nat Rev Mater* 2022, **7**:991-1009.

59. König-Mattern L, Medina EIS, Komarova AO, Linke S, Rihko-Struckmann L, Luterbacher JS, Sundmacher K: **Machine learning-supported solvent design for lignin-first biorefineries and lignin upgrading**. *Chem Eng J* 2024, **495**:153524.

60. Brozos C, Rittig JG, Akanny E, Bhattacharya S, Kohlmann C, Mitsos A: **Predicting the temperature-dependent CMC of surfactant mixtures with graph neural networks**. *Comput Chem Eng* 2025, **198**:109085.

61. Qin S, Jiang S, Li J, Balaprakash P, Lehn RCV, Zavala VM: **Capturing molecular interactions in graph neural networks: a case study in multi-component phase equilibrium**. *Digit Discov* 2023, **2**:138-151.

62. Rittig JG, Felton KC, Lapkin AA, Mitsos A: **Gibbs-Duhem-informed neural networks for binary activity coefficient prediction**. *Digit Discov* 2023, **2**:1752-1767.

63. E.M. Rajaonson, M.R. Kochi, L.M.M. Mendoza, S.M. Moosavi, B. Sanchez-Lengeling: **Chemixhub: Datasets and Benchmarks for Chemical Mixture Property Prediction**, arXiv preprint arXiv:2506.12231. https://doi.org/10.48550/arXiv.2506.12231.

64. A. Wahyudi, N. Sueviriyapan, T. Rirksomboon, U. Suriyapraphadilok, et al.: **Deepthermomix: A Local Composition Graph Neural Networks Model for Multicomponent Activity Coefficients**, chemRxiv preprint chemrxiv.10001495/v1. https://doi.org/10.26434/chemrxiv.10001495/v1.

65. Jirasek F, Hasse H: **Machine learning of thermophysical properties**. *Fluid Phase Equilibria* 2021, **549**:113206.

66. Chen G, Song Z, Qi Z, Sundmacher K: **Neural recommender system for the activity coefficient prediction and UNIFAC model extension of ionic liquid–solute systems**. *AIChE J* 2021, **67**:e17171.

67. SanchezMedina EI, Linke S, Stoll M, Sundmacher K: **Graph neural networks for the prediction of infinite dilution activity coefficients**. *Digit Discov* 2022, **1**:216-225.

68. Winter B, Winter C, Esper T, Schilling J, Bardow A: **SPT-NRTL: a physics-guided machine learning model to predict thermodynamically consistent activity coefficients**. *Fluid Phase Equilibria* 2023, **568**:113731.

69. M. Hoffmann, T. Specht, Q. GÃ¸ttl, J. Burger, S. Mandt, H. Hasse, F. Jirasek: **Thermodynamically consistent machine learning model for excess Gibbs energy**, arXiv preprint arXiv:2509.06484. https://doi.org/10.48550/arXiv.2509.06484.

70. Tom G, Ser CT, Rajaonson EM, Lo S, Park HS, Lee BK, Sanchez-Lengeling B: **Does this smell the same? Learning representations of olfactory mixtures using inductive biases**. *Mach Learn Sci Technol* 2025, **6**:035063.

71. S. Chithrananda, G. Grand, B. Ramsundar: **ChemBERTa: Large-scale Self-supervised Pretraining for Molecular Property Prediction**, arXiv preprint arXiv:2010.09885. https://doi.org/10.48550/arXiv.2010.09885.

72. Jiang D, Wu Z, Hsieh C-Y, Chen G, Liao B, Wang Z, Shen C, Cao D, Wu J, Hou T: **Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models**. *J Chemin* 2021, **13**:12.

73. J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M.M.A. Patwary, Y. Yang, Y. Zhou: **Deep Learning Scaling Is Predictable, Empirically**, arXiv preprint arXiv:1712.00409. https://doi.org/10.48550/arXiv.1712.00409.

74. J. Kaplan, S. McCandlish, T. Henighan, T.B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei: **Scaling Laws for Neural Language Models**, arXiv preprint arXiv:2001.08361. https://doi.org/10.48550/arXiv.2001.08361.

75. J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D.d.L. Casas, L.A. Hendricks, J. Welbl, A. Clark, et al.: **Training Compute-optimal Large Language Models**, arXiv preprint arXiv:2203.15556. https://doi.org/10.48550/arXiv.2203.15556.

76. J.G. Rittig: **Graph Machine Learning for Molecular Property Prediction and Design**, Ph.D. thesis, Dissertation, Rheinisch-Westfälische Technische Hochschule Aachen; 2025. https://doi.org/10.18154/RWTH-2025-04861.

77. K. Adams, L. Pattanaik, C.W. Coley: **Learning 3D Representations of Molecular Chirality with Invariance to Bond Rotations**, arXiv preprint arXiv:2110.04383. https://doi.org/10.48550/arXiv.2110.04383.

78. Schweidtmann AM, Rittig JG, Weber JM, Grohe M, Dahmen M, Leonhard K, Mitsos A: **Physical pooling functions in graph neural networks for molecular property prediction**. *Comput Chem Eng* 2023, **172**:108202.

79. Jirasek F, Hasse H: **Combining machine learning with physical knowledge in thermodynamic modeling of fluid mixtures**. *Annu Rev Chem Biomol Eng* 2023, **14**:31-51.

80. Felton KC, Raßpe-Lange L, Rittig JG, Leonhard K, Mitsos A, Meyer-Kirschner J, Knösche C, Lapkin AA: **ML-SAFT: a machine learning framework for PCP-SAFT parameter prediction**. *Chem Eng J* 2024,151999.

81. Habicht J, Brandenbusch C, Sadowski G: **Predicting PC-SAFT pure-component parameters by machine learning using a molecular fingerprint as key input**. *Fluid Phase Equilibria* 2023, **565**:113657.

82. Rosenberger D, Barros K, Germann TC, Lubbers N: **Machine learning of consistent thermodynamic models using automatic differentiation**. *Phys Rev E* 2022, **105**:045301.

83. G. Lastrucci, A.M. Schweidtmann: **ENFORCE: Nonlinear Constrained Learning with Adaptive-depth Neural projection**, arXiv preprint arXiv:2502.06774. https://doi.org/10.48550/arXiv.2502.06774.

84. Iftakher A, Golder R, Roy BN, Hasan MMF: **Physics-informed neural networks with hard nonlinear equality and inequality constraints**. *Comput Chem Eng* 2025,109418, https://doi.org/10.1016/j.compchemeng.2025.109418 Elsevier.

85. •• Chaparro G, Müller EA: **Development of a Helmholtz free energy equation of state for fluid and solid phases via artificial neural networks**. *Commun Phys* 2024, **7**:406.
This work proposes a machine learning approach that combines neural networks with the Helmholtz free energy to learn fluid-solid equation of states from property data with thermodynamic consistency.

86. Ramos MC, Collison CJ, White AD: **A review of large language models and autonomous agents in chemistry**. *Chem Sci* (6) 2025, **16**:2514-2572.

87. A. Mirza, N. Alampara, M. Ríos-García, M. Abdelalim, J. Butler, B. Connolly, T. Dogan, M. Nezhurina, B. Şen, S. Tirunagari, et al.: **ChemPile: A 250GB Diverse and Curated Dataset for Chemical Foundation Models**, arXiv preprint arXiv:2505.12534. https://doi.org/10.48550/arXiv.2505.12534.

88. Dutta S, Leal de Freitas I, MacielXavier P, Miceli de Farias C, BernalNeira DE: **Federated learning in chemical engineering: a tutorial on a framework for privacy-preserving collaboration across distributed data sources**. *Ind Eng Chem Res* 2025, **64**:7767-7783.

89. •• J.G. Rittig, C. Kortmann: **Federated Learning from Molecules to Processes: A Perspective**. arXiv preprint arXiv:2506.18525. https://doi.org/10.48550/arXiv.2506.18525.
This work provides a perspective on collaborative, data privacy-preserving training of machine learning models in the chemical engineering domain, enabling utilization of proprietary data from chemical companies and databank organizations.

90. Ramakrishnan R, Dral PO, Rupp M, von Lilienfeld OA: **Quantum chemistry structures and properties of 134 kilo molecules**. *Sci Data* 2014, **1**:140022.

91. Ruddigkeit L, van Deursen R, Blum LC, Reymond J-L: **Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17**. *J Chem Inf Model* 2012, **52**:2864-2875.

92. R. Bommasani, D.A. Hudson, E. Adeli, R.B. Altman, S. Arora, S. vonArx, M.S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N.S. Chatterji, A.S. Chen, K. Creel, J.Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L.E. Gillespie, K. Goel, N.D. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P.W. Koh, M.S. Krass, R. Krishna, R. Kuditipudi, et al.: **On the Opportunities and Risks of Foundation Models**. arXiv preprint arXiv:2108.07258. https://doi.org/10.48550/arXiv.2108.07258.

93. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M, *et al*.: **Analyzing learned molecular representations for property prediction**. *J Chem Inf Model* 2019, **59**:3370-3388.

94. J. Burns, A. Zalte, W. Green: **Descriptor-based Foundation Models for Molecular Property Prediction**. arXiv preprint arXiv:2506.15792. https://doi.org/10.48550/arXiv.2506.15792.

95. binJavaid M, Gervens T, Mitsos A, Grohe M, Rittig JG: **Exploring data augmentation: multi-task methods for molecular property prediction**. *Comput Chem Eng* 2025,109253.

96. Eraqi BA, Khizbullin D, Nagaraja SS, Sarathy SM: **Molecular property prediction in the ultra-low data regime**. *Commun Chem* 2025, **8**:201.

97. D. Beaini, S. Huang, J.A. Cunha, Z. Li, G. Moisescu-Pareja, O. Dymov, S. Maddrell-Mander, C. McLean, F. Wenkel, L. Müller, et al.: **Towards Foundational Models for Molecular Learning on Large-scale Multi-task Datasets.** arXiv preprint arXiv:2310.04292. https://doi.org/10.48550/arXiv.2310.04292.

98. K. Kläser, B. Banaszewski, S. Maddrell-Mander, C. McLean, L. Müller, A. Parviz, S. Huang, A. Fitzgibbon: **MiniMol: A Parameter-efficient Foundation Model for Molecular Learning**. arXiv preprint arXiv:2404.14986. https://doi.org/10.48550/arXiv.2404.14986.

99. Rodríguez-Pérez R, Bajorath J: **Explainable machine learning for property predictions in compound optimization**. *J Med Chem* 2021, **64**:17744-17752.

100. •• Wellawatte GP, Gandhi HA, Seshadri A, White AD: **A perspective on explanations of molecular prediction models**. *J Chem Theory Comput* 2023, **19**:2149-2160.
This work provides a comprehensive overview and perspective on explainability methods for molecular machine learning models.

101. Yuan H, Yu H, Gui S, Ji S: **Explainability in graph neural networks: a taxonomic survey**. *IEEE Trans Pattern Anal Mach Intell* 2023, **45**:5782-5799.

102. •• Teufel J, Friederich P: **Global concept explanations for graphs by contrastive learning**. World Conference on Explainable Artificial Intelligence. Springer; 2024:184-208.
This work proposes an approach for model-level explanations of GNNs, which enables uncovering structure-property relationships from molecular property data.

103. • Gond D, Sohns J-T, Leitte H, Hasse H, Jirasek F: **Hierarchical matrix completion for the prediction of properties of binary mixtures**. *Comput Chem Eng* 2025,109122.
This work combines matrix completion and hierarchical clustering methods to reveal chemical insights at the molecular structural level that are crucial for predicting a specific mixture property.

104. Ignatiev A, Narodytska N, Asher N, Marques-Silva J: **From contrastive to abductive explanations and back again**. International Conference of the Italian Association for Artificial Intelligence. Springer; 2020:335-355.

105. Wellawatte GP, Seshadri A, White AD: **Model agnostic generation of counterfactual explanations for molecules**. *Chem Sci* 2022, **13**:3697-3705.

106. Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW: **Uncertainty quantification using neural networks for molecular property prediction**. *J Chem Inf Model* 2020, **60**:3770-3780.

107. Jiang S, Qin S, Van Lehn RC, Balaprakash P, Zavala VM: **Uncertainty quantification for molecular property predictions with graph neural architecture search**. *Digit Discov* 2024, **3**:1534-1553.

108. Gaoa Q, Miedemaa DC, Zhaob Y, Weberc JM, Taob Q, Schweidtmanna AM: **Bayesian uncertainty quantification of graph neural networks using stochastic gradient Hamiltonian Monte Carlo**. *Syst Control Trans* 2025,1360-1364.

109. •• Heid E, McGill CJ, Vermeire FH, Green WH: **Characterizing uncertainty in machine learning for chemistry**. *J Chem Inf Model* 2023, **63**:4012-4029.
This work analyzes sources of uncertainty in property data and molecular machine learning models, providing guidelines and research

directions for uncertainty quantification in the property prediction context.

110. Komissarov L, Manevski N, GroebkeZbinden K, Sach-Peltason L: **Explainable graph neural networks in chemistry: Combining attribution and uncertainty quantification**. *J Chem Inf Model* 2025, **65**:7516-7528.

111. J. Teufel, A. Leinweber, P. Friederich: **Improving Counterfactual Truthfulness for Molecular Property Prediction Through Uncertainty Quantification.** arXiv preprint arXiv:2504.02606. doi: 10.48550/arXiv.2504.02606.

112. Mann V, Gani R, Venkatasubramanian V: **Group contribution-based property modeling for chemical product design: a perspective in the AI era**. *Fluid Phase Equilibria* 2023, **568**:113734.

113. Rittig JG, Ritzert M, Schweidtmann AM, Winkler S, Weber JM, Morsch P, Heufer KA, Grohe M, Mitsos A, Dahmen M: **Graph machine learning for design of high–octane fuels**. *AIChE J* 2023, **69**:e17971.

114. Sarathy SM, Eraqi BA: **Artificial intelligence for novel fuel design**. *Proc Combust Inst* 2024, **40**:105630.

115. Vogel G, Weber JM: **Inverse design of copolymers including stoichiometry and chain architecture**. *Chem Sci* 2025, **16**:1161-1178.

116. Pirnay J, Rittig JG, Wolf AB, Grohe M, Burger J, Mitsos A, Grimm DG: **GraphXForm: graph transformer for computer-aided molecular design**. *Digit Discov* 2025, **4**:1052-1065, https://doi.org/10.1039/D4DD00339

117. Iftakher A, Hasan MF: **Design space exploration and machine learning prediction of hydrofluorocarbon solubility in ionic liquids for refrigerant separation**. *J Chem Inf Model* 2025, **65**:12168-12178.

118. • J. Guo, V. Sabanza-Gil, Z. Jončev, J.S. Luterbacher, P. Schwaller: **Generative molecular design with steerable and granular synthesizability control.** arXiv preprint arXiv:2505.08774..
This work proposes a generative machine learning model that enables the design of molecules with desired properties from specific building blocks and accounts for synthesizability.

119. Tu Z, Choure SJ, Fong MH, Roh J, Levin I, Yu K, Joung JF, Morgan N, Li S-C, Sun X, *et al.*: **ASKCOS: open-source, data-driven synthesis planning**. *Acc Chem Res* 2025, **58**:1764-1775.

120. • McDonald T, Tsay C, Schweidtmann AM, Yorke-Smith N: **Mixed-integer optimisation of graph neural networks for computer-aided molecular design**. *Comput Chem Eng* 2024, **185**:108660.
This work proposes deterministic global optimization of GNNs embedded into optimization problems with an application to molecular design.

121. Zhang S, Campos J, Feldmann C, Walz D, Sandfort F, Mathea M, Tsay C, Misener R: **Optimizing over trained GNNs via symmetry breaking**. In *Advances in Neural Information Processing Systems*. Edited by Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S. 36 Curran Associates, Inc.; 2023:44898-44924 https://proceedings.neurips.cc/paper_files/paper/2023/file/8c8cd1b78cdae751265c88efc136e5bd-Paper-Conference.pdf.

122. Rittig JG, Franke M, Mitsos A: **Deterministic global optimization for sample-efficient molecular design with generative machine learning**. *AI for Accelerated Materials Design-NeurIPS 2024*. 2024.

123. Wang Z, Zhou T, Sundmacher K: **A novel machine learning-based optimization approach for the molecular design of solvents**. Computer Aided Chemical Engineering Elsevier; 2022:1477-1482.

124. Scheffczyk J, Fleitmann L, Schwarz A, Lampe M, Bardow A, Leonhard K: **COSMO-CAMD: a framework for optimization-based computer-aided molecular design using COSMO-RS**. *Chem Eng Sci* 2017, **159**:84-92.

125. •• Adjiman CS, Galindo A: **Challenges and opportunities for computer-aided molecular and process design approaches in advancing sustainable pharmaceutical manufacturing**. *Curr Opin Chem Eng* 2025, **47**:101073.
This opinion article discusses recent and future research on integrated molecular and process design.

126. Iftakher A, Monjur MS, Hasan MF: **An overview of computer-aided molecular and process design**. *Chem Ing Tech* 2023, **95**:315-333.

127. Rehner P, Schilling J, Bardow A: **Molecule superstructures for computer-aided molecular and process design**. *Mol Syst Des Eng* 2023, **8**:488-499.

128. • Bosetti L, Winter B, Lindfeld J, Bardow A: **Integrated design of solvent-antisolvent mixtures and crystallization processes powered by machine learning**. *Comput Chem Eng* 2025,109272.
This work integrates molecular machine learning with the process scale to enable the simultaneous design of solvent-antisolvent mixtures and optimization of crystallization processes.

129. Ceccon F, Jalving J, Haddad J, Thebelt A, Tsay C, Laird CD, Misener R: **OMLT: optimization & machine learning toolkit**. *J Mach Learn Res* 2022, **23**:1-8.

130. A.M. Schweidtmann, L. Netze, A. Mitsos, **MeLOn - Machine Learning Models for Optimization**; 2021.⟨https://git.rwth-aachen.de/avt-svt/public/MeLOn⟩ (accessed on 01.08.2025).

131. Wang Z, Zhou T, Sundmacher K: **Bayescampd: data-efficient and closed-loop integrated molecular and process design using bayesian optimization**. *AIChE J* 2025,e70191.

132. •• Schweidtmann AM: **Generative artificial intelligence in chemical engineering**. *Nat Chem Eng* 2024, **1**:193.
This work provides a short review and opinion on the use of generative machine learning for chemical process design.

133. Gao Q, Schweidtmann AM: **Deep reinforcement learning for process design: review and perspective**. *Curr Opin Chem Eng* 2024, **44**:101012.

134. • Göttl Q, Pirnay J, Burger J, Grimm DG: **Deep reinforcement learning enables conceptual design of processes for separating azeotropic mixtures without prior knowledge**. *Comput Chem Eng* 2025, **194**:108975.
This work proposes a reinforcement framework for the design of flowsheets for separation processes that enables generalization to feeds with different chemical species.

135. Rupprecht S, Gao Q, Karia T, Schweidtmann AM: **Multi-agent systems for chemical engineering: a review and perspective**. *Curr Opin Chem Eng* 2026, **51**:101209, https://doi.org/10.1016/j.coche.2025.101209 Elsevier.

136. Y. Du, B. Yu, T. Liu, T. Shen, J. Chen, J.G. Rittig, K. Sun, Y. Zhang, Z. Song, B. Zhou, et al.: **Accelerating Scientific Discovery with Autonomous Goal-evolving Agents.** arXiv preprint arXiv:2512.21782. https://doi.org/10.48550/arXiv.2512.21782.