REPRESENTING SURFACTANTS BY FOUNDATION MODELS

Eduardo. Soares, Zeynep Sumer, Emilio Vital Brazil & Dave Braines
IBM Research
{eduardo.soares,zsumer}@ibm.com
{evital}@br.ibm.com

Richard L. Anderson
Hartree Centre STFC Laboratory Sci-Tech

ABSTRACT

This work presents a novel approach to predicting surfactant phase diagrams by leveraging the SMI-TED$_{289M}$ foundation model, a pre-trained encoder-decoder architecture based on SMILES representations. The methodology integrates molecular representations with environmental variables, including composition (wt%) and temperature (°C), to enhance predictive performance. For phase diagram prediction, the latent space of SMI-TED was extended with thermodynamic parameters. Experimental results demonstrate accurate predictions for dominant phases such as liquid, ice and aqueous phases, with phase boundaries closely aligned with experimental data. However, the model exhibits limitations in boundary and transition regions, particularly for minority phases like lamellar, cubic and solid surfactant phases. These findings highlight the potential of integrating molecular and thermodynamic data within foundation models for predictive materials science, while also identifying opportunities for improvement through enhanced data representation, thermodynamic constraints, and uncertainty quantification.

1 INTRODUCTION

Surfactants are amphiphilic molecules that are composed of two parts, hydrophilic (water-attracted) heads and hydrophobic (oil-attracted) tails. This polarization enables them to decrease the surface tension between two immiscible compounds, and to create homogeneous mixtures that is otherwise impossible to keep stable. The variety of applications include, but are not limited to medicines, e.g., treating respiratory distress syndrome,(1) energy industry where surfactants are used as fuel oil additives to reduce the gas evaporation,(2) cleaning and beauty, e.g., substrate removal from textile or ceramics and formulations,(3) and farming where surfactants are used for better attachment of pesticides on the leaf.(4)

Rapid development in computational tools eased the task of discovery of new materials and reduced the experimental demand, so surfactant science was also improved by using these techniques. For example, molecular simulations were adopted to calculate properties such as the interfacial tension, or critical micelle concentration (CMC) of surfactants in liquid medium.(5; 6; 7; 8; 9) Quantitative structure property/activity relationship (QSPR or QSAR) studies also escalated the improvement of surfactant property predictions.(10; 11; 12; 13) Lately, machine learning tools were being developed following the success of molecular simulations and QSPR/QSAR studies, and often integrated to these methods. (14; 15; 16)

In all computational study tasks, however, the main challenge lays in the representation of surfactants. In molecular simulations, classical force-fields often come with many assumptions that reliable set of parameters for one type of surfactant can fail in another (particularly with ionic surfactants). Moreover, representing the liquid medium is another challenge, e.g., representing water molecules is a widely known problem that could not yet be fully solved.(17) Even when the reliable force-fields are available, the molecular simulations come with a vast computational cost due to the large system sizes that should be analyzed within a time frame for the self-assembly studies. Adding few surfactants to

a water medium can only elucidate the dilute conditions but for phases such as laminar, hexagonal or micellar, ratio of surfactants to water molecules must increase, meaning number of iterative calculations increase.

Surfactant phase diagrams (considering the binary water-surfactant mixtures), reveal thermodynamic information about the self-assembly in the mixture at a given temperature and concentration. As certain thermodynamic phases are desired in a mixture for the target use (e.g., cold water cleaning, or micellization), phase diagrams show whether it is possible to make an efficient product with the surfactant. Moreover, phase diagrams can elucidate many properties such as cloud point, CMC and so forth. Thacker et al.(18) and Sumer et al.(19) analyzed several machine learning models for the accurate prediction of phase diagrams, but conclusion in both works underlined the importance of the representation of materials, i.e., descriptors, alongside the effective representation of phase data, which was the main challenge in another work. (20)

In this work we evaluated the capacity of the foundation model, SMI-TED, (21) in describing various surfactants, and built predictive models to discover their phase diagrams within certain temperature and concentration ranges in water/surfactant mixtures. We addressed the chemical representation challenges and results revealed that not only we achieved the highest accuracy in predicting phase diagrams so far, but also the foundation model used in this work can increase the capacity of the models for e.g., any other surfactant/solvent mixture, or ternary mixtures, which are industrially more relevant compositions; not only for phase prediction, but also for other important properties e.g., CMC, interfacial tension.

2 Overview of the Proposed Approach

This section provides an overview of the SMI-TED$_{289M}$ foundation model and its adaptation for predicting surfactant phase diagrams. The methodology begins by describing the processes involved in collecting, curating, and pre-processing molecular data for pre-training. It then explains the encoder-decoder architecture of the SMI-TED$_{289M}$ model, which is pre-trained on SMILES representations to learn compact and expressive latent molecular embeddings. Finally, the section introduces the extension of the model to incorporate additional environmental variables, such as composition (wt%) and temperature (°C), to improve its performance in predicting phase behavior.

2.1 Model Architecture

To evaluate the capacity of foundation models to predict surfactants phase diagrams, we utilized the SMI-TED$_{289M}$ foundation model as SMILES encoder. (21) SMI-TED$_{289M}$ is an open-source encoder-decoder model pre-trained on a curated dataset of 91 million SMILES samples from PubChem. This model has demonstrated superior performance compared to state-of-the-art methods across various molecular tasks. (21)

The molecular tokenizer proposed by Schwaller and coworkers was employed to construct the vocabulary of SMI-TED$_{289M}$.(22) All 91 million molecules curated from PubChem were utilized in the tokenization process, resulting in a set of 4 billion molecular tokens. The unique tokens extracted from the resulting output provided a vocabulary of 2988 tokens plus 5 special tokens.

Pre-training of of SMI-TED$_{289M}$ was performed for 40 epochs through the entire curated PubChem dataset with a fixed learning rate of 1.6e-4 and a batch size of 288 molecules on a total of 24 NVIDIA V100 (16G) GPUs parallelized into 4 nodes using DDP and *torch run*. It involved two distinct phases: i) Learning of token embeddings through a masking process; ii) Subsequently, the token embeddings were mapped into a common latent space that encapsulates the entire SMILES string.

2.2 Model Adaptation for Phase Diagram Prediction

To adapt SMI-TED$_{289M}$ for surfactant phase diagram prediction, the model's latent space was extended to incorporate composition (wt%) and temperature (°C) as additional input features, as illustrated in Fig. 1. These features are important for capturing the thermodynamic dependencies of phase behavior, enabling the model to account for environmental conditions that significantly influence phase transitions. By integrating molecular representations from SMILES with environmental

parameters, the adapted model can learn the relationships between molecular structures and their macroscopic phase behavior.

The extension of the latent space involves appending the thermodynamic input features directly to the embeddings generated by the pre-trained SMI-TED$_{289M}$ model. This enriched representation provides a more comprehensive view of the factors influencing phase behavior, allowing the model to predict phase boundaries and transitions with improved accuracy. For downstream learning, an XGBoost model was employed, leveraging its flexibility and robustness in handling complex, high-dimensional data. To optimize predictive performance, hyperparameter tuning was conducted using Optuna, an optimization framework, ensuring the model parameters were tailored to the specific requirements of phase diagram prediction.

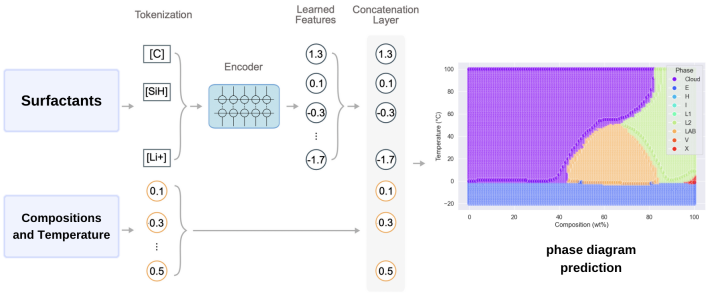

Figure 1: Adaptation of SMI-TED$_{289M}$ for phase diagram prediction, incorporating composition and temperature as additional input features alongside molecular embeddings derived from SMILES.

The proposed approach was evaluated on experimentally determined surfactant phase diagrams, using performance metrics such as confusion matrices to quantify predictive ability. As shown in the results, the adapted model demonstrated strong agreement with experimental data for dominant phases, including $Cloud$, ice (E), and aqueous (L_1) phases. These results highlight the model's ability to generalize thermodynamic trends effectively, capturing stable phase behavior across a wide range of conditions.

3 Experimental Data

We used a database of 65 non-ionic surfactants and their binary water/surfactant phase diagrams.(20) 10 of these compounds were used as the test set materials and the model was trained with the information from remaining 55 diagrams.(19) Database consists of set of surfactants that include alcohol ethoxylates, methoxypolyoxyethylenes, lipids, rigid amphiphiles, polyfluoroalkyl (PFA) surfactants and branched surfactants. (19)

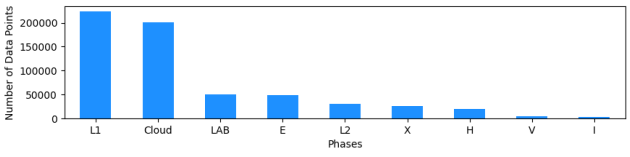

Figure 2: Number of data points for each phase in the database for total of 65 surfactants.

Each phase diagram of these surfactants consists of sample points across a temperature – surfactant concentration grid at intervals of 1 °C and 1 wt %. There were in total of 54 phase labels across the data set of 65 diagrams, with a strong class imbalance in the phase labels. We amalgamated these 54 phases into 9 phases to improve data distribution. Therefore the collection of diagrams contained a total of 9 unique surfactant phases. These phases include: the aqueous phase (L_1), the alcohol phase (or inverse micelles, L_2), hexagonal phase (H), cubic-bicontinuous phase (V), cubic-spherical-micelles (I), lamellar phase ($L_{\alpha\beta}$), ice phase (E), solid surfactant phase (X) and

finally two-phase regions with a coexisting liquid phase (W), which was referred as the $Cloud$ phase. Fig. 2, shows the number of data points per each class (phase) in the database.

4 Results and Discussion

The phase diagrams and confusion matrix provide a comprehensive evaluation of the predictive performance of the SMI-TED foundation model for surfactant systems. By incorporating composition (wt%) and temperature (°C) as additional input features alongside SMILES-based molecular representations, the model demonstrates its capability to predict phase behavior across diverse thermodynamic conditions, bridging molecular-scale features and macroscopic phase behavior.

As illustrated in Fig. 3, the SMI-TED model effectively reconstructs global phase behavior, particularly in well-defined regions dominated by single phases such as the $Cloud$, ice (E), and aqueous (L_1) phases. These dominant phases exhibit alignment between predicted and experimental phase diagrams, reflecting the model's ability to generalize thermodynamic trends. This capability is quantitatively supported by the confusion matrix in Fig. 4, which highlights high prediction accuracy for $Cloud$ and L_1, with 31,204 and 23,434 correct predictions, respectively. Other dominant phases, such as ice (E) (8,343 correct predictions) and inverse micelle (L_2) (5,537 correct predictions), further demonstrate the robustness of the model in capturing thermodynamically stable regions. High-temperature regions (above 80 °C) exhibit close agreement between predicted and experimental phase boundaries, indicating that the model is adept at capturing simpler phase transitions under such conditions.

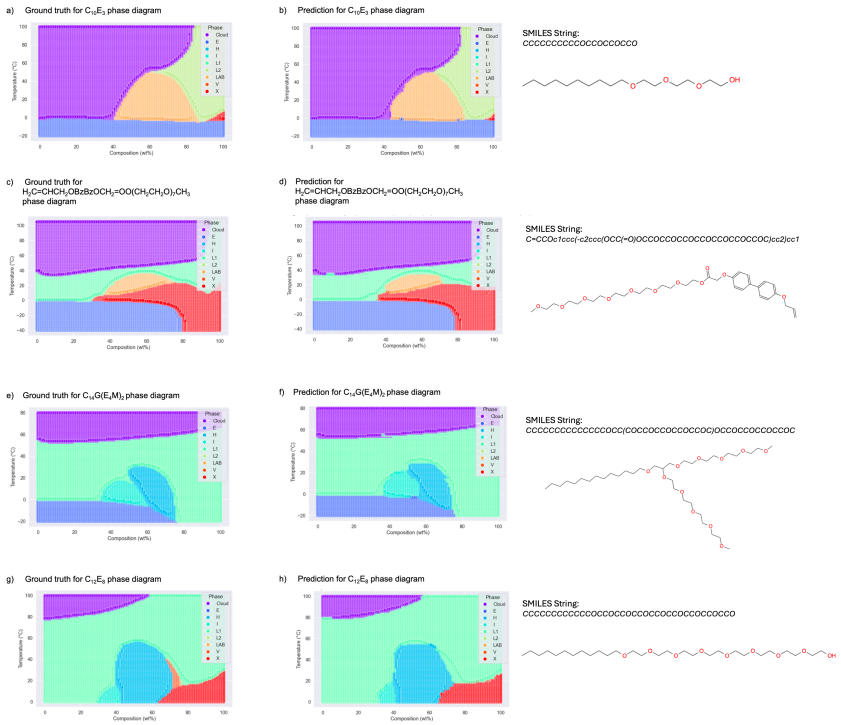

Figure 3: Comparison of ground truth and SMI-TED predicted phase diagrams for different surfactants. Subfigures (a), (c), (e), and (g) represent the ground truth phase diagrams for four randomly selected surfactants, illustrating the experimentally determined phase behavior across varying compositions and temperatures. Subfigures (b), (d), (f), and (h) depict the corresponding phase predictions generated by the SMI-TED model, highlighting its ability to approximate phase behavior under similar conditions.

The overall classification metrics further validate the model's strong performance. The macro-averaged precision, recall, and F1-score are 0.947, 0.935, and 0.941, respectively, demonstrating balanced performance across all classes. Moreover, the weighted averages for precision (0.972),

recall (0.973), and F1-score (0.972) indicate that the model performs particularly well for classes with higher representation, such as $Cloud$ and L_1, while accounting for the imbalance in class distributions.

However, despite these strengths, the model exhibits significant limitations when predicting boundary and minority phases. As shown in Fig. 4, phases such as hexagonal (H), cubic-spherical-micelles (I), cubic-bicontinuous (V), and solid (X) show lower prediction accuracy. For instance, H and I account for only 2,346 and 890 correct predictions, respectively, with frequent misclassifications into neighboring phases. Similarly, V and X are often misclassified, with 97 instances of V predicted as $Cloud$ and 58 instances of X predicted as the ice phase (E). These errors likely arise from overlapping phase features and the insufficient representation of these minority phases in the training dataset, leading to challenges in resolving molecular interactions and phase boundaries.

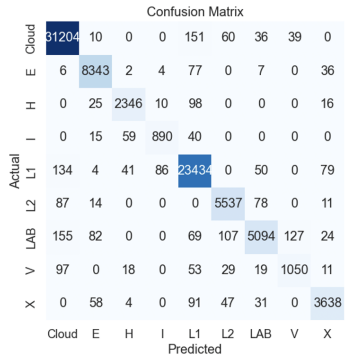

Figure 4: Confusion matrix for phase diagram predictions considering the entire test set.

The limitations observed in the predictions suggest several avenues for improvement. Addressing data imbalance through targeted augmentation and the inclusion of additional experimental data for minority phases is a critical step. Expanding the training dataset with examples from underrepresented regions and transition zones would help the model learn the nuanced features associated with these phases. Incorporating thermodynamic descriptors, such as free energy, enthalpy, or molecular interaction parameters, could further enhance the model's ability to differentiate between overlapping or minority phases by providing richer contextual information. Additionally, integrating uncertainty quantification mechanisms into the model could systematically identify regions where the model lacks confidence, enabling targeted experimental validation and data enrichment.

5 Conclusion

This work presents an integrated approach for predicting surfactant phase diagrams using the SMI-TED$_{289M}$ foundation model, which combines molecular representations with environmental variables such as composition (wt%) and temperature (°C). By leveraging the pre-trained capabilities of the SMI-TED$_{289M}$ model, the proposed methodology encodes SMILES strings into compact and expressive latent embeddings while extending the model to include thermodynamic parameters. This integration highlights the potential of foundation models to bridge the gap between molecular-scale features and macroscopic phase behavior, enabling more accurate predictions of complex phase behavior.

The results demonstrate that the SMI-TED$_{289M}$ model achieves strong predictive performance for dominant phases, such as $Cloud$, ice (E), and aqueous (L_1) phases, with phase boundaries closely aligned with experimental data. The confusion matrix quantitatively validates the model's robustness in capturing global phase trends, particularly for well-represented classes. However, it also reveals limitations in predicting boundary and minority phases, such as lamellar ($L_{\alpha\beta}$), cubic (V), and solid (X) phases. These limitations are attributed to the increased complexity of overlapping phase interactions and the under-representation of these classes in the training dataset. To overcome current limitations, the model could be improved by augmenting the training data, incorporating thermodynamic descriptors, and integrating domain-specific constraints to enhance generalization, feature richness, and physical plausibility of predictions.

REFERENCES

[1] Smeeta Sardesai, Manoj Biniwale, Fiona Wertheimer, Arlene Garingo, and Rangasamy Ramanathan. Evolution of surfactant therapy for respiratory distress syndrome: past, present, and future. *Pediatric research*, 81(1):240–248, 2017.

[2] Jinesh Machale, Duraid Al-Bayati, Mohamed Almobarak, Mohsen Ghasemi, Ali Saeedi, Tushar Kanti Sen, Subrata Kumar Majumder, and Pallab Ghosh. Interfacial, emulsifying, and rheological properties of an additive of a natural surfactant and polymer and its performance assessment for application in enhanced oil recovery. *Energy & Fuels*, 35(6):4823–4834, 2021.

[3] Gunjan Tyagi, Zain Ahmad, Luca Pellegrino, Luis MG Torquato, Eric SJ Robles, and João T Cabral. Effect of surface energy on the removal of supported triglyceride films by a flowing surfactant solution. *Surfaces and Interfaces*, 39:102992, 2023.

[4] Mustafa O Jibrin, Qingchun Liu, Jeffrey B Jones, and Shouan Zhang. Surfactants in plant disease management: A brief review and case studies. *Plant Pathology*, 70(3):495–510, 2021.

[5] Takeshi Kobayashi, Kristo Kotsi, Teng Dong, Ian McRobbie, Alexander Moriarty, Panagiota Angeli, and Alberto Striolo. The solvation of na+ ions by ethoxylate moieties enhances adsorption of sulfonate surfactants at the air-water interface. *Journal of Colloid and Interface Science*, 682:924–933, 2025.

[6] Tseden Taddese, Richard L. Anderson, David J. Bray, and Patrick B. Warren. Recent advances in particle-based simulation of surfactants. *Current Opinion in Colloid Interface Science*, 48:137–148, 2020. Formulations and Cosmetics.

[7] Michael A. Johnston, Andrew Ian Duff, Richard L. Anderson, and William C. Swope. Model for the simulation of the cnem nonionic surfactant family derived from recent experimental results. *The Journal of Physical Chemistry B*, 124(43):9701–9721, 2020. PMID: 32986421.

[8] Harry Cárdenas, M. Ariif H. Kamrul-Bahrin, Dale Seddon, Jofry Othman, João T. Cabral, Andrés Mejía, Sara Shahruddin, Omar K. Matar, and Erich A. Müller. Determining interfacial tension and critical micelle concentrations of surfactants from atomistic molecular simulations. *Journal of Colloid and Interface Science*, 674:1071–1082, 2024.

[9] Paulina Müller, Douwe Jan Bonthuis, Reinhard Miller, and Emanuel Schneck. Ionic surfactants at air/water and oil/water interfaces: A comparison based on molecular dynamics simulations. *The Journal of Physical Chemistry B*, 125(1):406–415, 2021. PMID: 33400514.

[10] Danial Abooali and Reza Soleimani. Structure-based modeling of critical micelle concentration (cmc) of anionic surfactants in brine using intelligent methods. *Scientific Reports*, 13(1):13361, 2023.

[11] Jiaqi Wu, Fangyou Yan, Qingzhu Jia, and Qiang Wang. Qspr for predicting the hydrophile-lipophile balance (hlb) of non-ionic surfactants. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 611:125812, 2021.

[12] Nada Boukelkal, Soufiane Rahal, Redha Rebhi, and Mabrouk Hamadache. Qspr for the prediction of critical micelle concentration of different classes of surfactants using machine learning algorithms. *Journal of Molecular Graphics and Modelling*, 129:108757, 2024.

[13] James Y Liu, Joshua Peeples, and Christie M Sayes. Evaluation of machine learning based qsar models for the classification of lung surfactant inhibitors. *Environment & Health*, 2(12):912–917, 2024.

[14] Dale Seddon, Erich A Müller, and João T Cabral. Machine learning hybrid approach for the prediction of surface tension profiles of hydrocarbon surfactants in aqueous solution. *Journal of Colloid and Interface Science*, 625:328–339, 2022.

[15] Christoforos Brozos, Jan G Rittig, Sandip Bhattacharya, Elie Akanny, Christina Kohlmann, and Alexander Mitsos. Graph neural networks for surfactant multi-property prediction. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 694:134133, 2024.

[16] Miriam Nnadili, Andrew N Okafor, Teslim Olayiwola, David Akinpelu, Revati Kumar, and Jose A Romagnoli. Surfactant-specific ai-driven molecular design: Integrating generative models, predictive modeling, and reinforcement learning for tailored surfactant synthesis. *Industrial & Engineering Chemistry Research*, 63(14):6313–6324, 2024.

[17] Sachini P Kadaoluwa Pathirannahalage, Nastaran Meftahi, Aaron Elbourne, Alessia CG Weiss, Chris F McConville, Agilio Padua, David A Winkler, Margarida Costa Gomes, Tamar L Greaves, Tu C Le, et al. Systematic comparison of the structural and dynamic properties of commonly used water models for molecular dynamics simulations. *Journal of Chemical Information and Modeling*, 61(9):4521–4536, 2021.

[18] Joseph CR Thacker, David J Bray, Patrick B Warren, and Richard L Anderson. Can machine learning predict the phase behavior of surfactants? *The Journal of Physical Chemistry B*, 127(16):3711–3727, 2023.

[19] Zeynep Sumer, Jonathan Booth, Patrick B. Warren, David J. Bray, Felix Rummel, Ardita Shkurti, and Richard L. Anderson. A multiclass classification model for systematic evaluation and prediction of surfactant phase diagrams. *In preparation*.

[20] Felix Rummel, Patrick B. Warren, David J. Bray, Zeynep Sumer, Jonathan Booth, Ardita Shkurti, and Richard L. Anderson. Surfactant phase behavior dataset. *In preparation*.

[21] Eduardo Soares, Victor Shirasuna, Emilio Vital Brazil, Renato Cerqueira, Dmitry Zubarev, and Kristin Schmidt. A large encoder-decoder family of foundation models for chemical language. *arXiv preprint arXiv:2407.20267*, 2024.

[22] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.

[23] JC Ravey and MJ Stébé. Properties of fluorinated non-ionic surfactant-based systems and comparison with non-fluorinated systems. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 84(1):11–31, 1994.

6 Supplementary Materials

Challenges of the model are further exemplified in Fig. 5, where the fluorinated surfactant $C_6F(EO)_5OME$ highlights the model's failure to predict minority phases.(23) This surfactant's phase diagram demonstrates complete omission of minority classes in the SMI-TED predictions, indicating the model's limited ability to generalize in underrepresented or boundary regions. Additionally, the mid-composition range (50–70 wt%) and moderate temperatures (30–60 °C) represent particularly challenging regions, with frequent misclassifications such as 127 instances of $L_{\alpha\beta}$ predicted as L_2 and 107 instances as V. These transition zones, characterized by intricate molecular interactions and equilibrium states, remain difficult for the current model to resolve with high confidence.

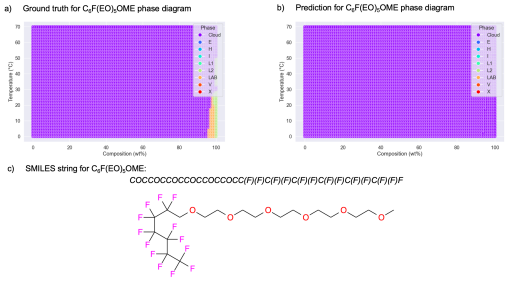

Figure 5: (a) Ground truth phase diagram for $C_6F(EO)_5OME$. (b) SMI-TED phase prediction, where minority classes are omitted. (c) SMILES string for $C_6F(EO)_5OME$ and its schematical representation.