A Framework for Toxic PFAS Replacement based on GFlowNet and Chemical Foundation Model

Eduardo Soares
IBM Research Brazil
Rio de Janeiro, RJ, Brazil
eduardo.soares@ibm.com

Flaviu Cipcigan
IBM Research Europe
Warrington, United Kingdom
flaviu.cipcigan@ibm.com

Dmitry Zubarev
IBM Research Almaden
San Jose, CA, USA
dmitry.zubarev@ibm.com

Emilio Vital Brazil
IBM Research Brazil
Rio de Janeiro, RJ, Brazil
evital@br.ibm.com

Abstract

Per- and polyfluoroalkyl substances (PFAS) are a broad class of molecules used in almost every sector of industry and consumer goods. PFAS exhibit highly desirable properties such as high durability, water repellance or high acidity, that are difficult to match. As a side effect, PFAS persist in the environment and have detrimental effect on human health. Epidemiological research has linked PFAS exposure to chronic health conditions, including dyslipidemia, cardiometabolic disorders, liver damage, and hypercholesterolemia. Recently, public health agencies significantly strengthed regulations on the use of PFAS. Therefore, alternatives are needed to maintain the pace of technological developments in multiple areas that traditionally relied on PFAS. To support the discovery of alternatives, we introduce MatGFN-PFAS, an AI system that generates PFAS replacements. We build MatGFN-PFAS using Generative Flow Networks (GFlowNets) for generation and a Chemical Language Model (MolFormer) for property prediction. We evaluate MatGFN-PFAS by exploring potential replacements of PFAS superacids, defined as molecules with negative pKa, that are critical for the semiconductor industry. It might be challenging to eliminate PFAS superacids entirely as a class due to the strong constraints on their functional performance. The proposed approach aims to account for this possibility and enables the generation of safer PFAS superacids as well. We evaluate two design strategies: 1) Using Tversky similarity to design molecules similar to a target PFAS and 2) Directly generating molecules with negative pKa and low toxicity. In this paper, we studied 6 PFAS molecules that have the structure defined as $R - CF_2OCF_2 - R'$. For the given query PFAS SMILE **CC1CC(CC(F)(F)C(F)(F)OC(F)(F)C(F)(F)S(=O)(=O)O)OC1=O**, MatGFN-PFAS system was able to generate a candidate with very low toxicity, $LD50 = 7304.23$, strong acidity, $pKa = -1.92$, and high similarity score, 89.32%, to the studied PFAS molecule. Results demonstrated that the proposed MatGFN-PFAS was able to consistently generate replacement molecules following all the constraints forehead mentioned. The resulting datasets for this ongoing study are available at https://ibm.box.com/v/MatGFN-PFAS-generated-datasets.

1 Introduction

Per- and polyfluoroalkyl substances (PFAS) belong to a category of compounds within the broader domain of organofluorine substances [1]. In 2021, the Organisation for Economic Cooperation

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

and Development (OECD) introduced an updated definition for PFAS [2], characterizing them as fluorinated substances that incorporate at least one fully fluorinated methyl or methylene carbon atom, devoid of any hydrogen (H), chlorine (Cl), bromine (Br), or iodine (I) atoms bound to it. PFAS have demonstrated their utility across a diverse spectrum of consumer goods and industrial applications [3].

While PFAS offer exceptional advantages in the production of both consumer and industrial items, their robustness and durability also contribute to their remarkable resistance to degradation [4]. Consequently, numerous older-generation PFAS have accumulated in the environment, wildlife, and human beings over time [5, 6, 7]. As a result, a blend of legacy PFAS, as well as newer alternatives and emerging variants, persistently pervades our surroundings, notably including global food and water supplies [8, 9]. Furthermore, the bioaccumulation of PFAS in human blood has been widely documented, with detection in approximately 98% of adult Americans [10].

Given the extensive bioaccumulation and persistence characteristics of PFAS, the field of epidemiological research focusing on the adverse health impacts of PFAS exposure in humans is rapidly expanding [11]. A multitude of epidemiological investigations has identified associations between PFAS exposure and chronic ailments such as dyslipidemia and cardiometabolic disorders [12, 13, 14, 15]. Human studies have also established connections between elevated serum PFAS levels and liver damage [16, 17], including increased alanine aminotransferase (ALT) levels [18], steatosis [19], and non-alcoholic fatty liver disease (NAFLD) severity [14]. Furthermore, epidemiological revealed links between PFAS exposure and elevated cholesterol and triglyceride levels [20, 21].

Stated the significant adverse impacts on both ecological and human health associated with PFAS substances, there is an urgent need for research aimed at identifying potential alternatives to mitigate these harmful effects [22]. This paper introduces MatGFN-PFAS, an AI system that harnesses the power of the recently developed Generative Flow Network (GFlowNet) [23]. In this particular study, we employ MoLFormer [24], a large chemical language model, to predict the toxicity (LD50) [25] and pKa of the generated molecules, which serves as the basis for the GFlowNet's reward function. Furthermore, we incorporate Tversky similarity within the reward function to generate molecules with a structural resemblance to the PFAS molecules currently under investigation [26].

2 Methodology

In this section, we explain the methodological framework of MatGFN-PFAS delineated within this study. As depicted in Figure 1, we present an intricately devised schema for the generation of candidates to replace PFAS substances leveraging GFlowNet. This approach takes into account three key components: the MoLFormer-derived toxicity and pKA prediction for the synthesized molecule and the Tversky similarity metric quantifying the resemblance between the target molecule (the entity to be replaced) and the synthesized molecule, weighing shared structures. These constituents are systematically employed as components of a reward function, aiming the generation of optimal candidate molecules for the PFAS replacement endeavor.

2.1 GFlowNet

The GFlowNet algorithm is designed to learn a generative policy that constructs objects by following a sequence of actions, as detailed in the comprehensive description provided by [23]. This policy operates within a user-defined deterministic Markov decision process (MDP) [27]. The MDP encompasses a state space denoted as S, a set of permissible actions A_s associated with each state s, a deterministic transition function $S \times A_s \to S$, and a reward function R. GFlowNets conceptualize this MDP as a directed graph referred to as a flow network. In this representation, states correspond to nodes, and the MDP's transition function defines directed edges between these nodes. A state's children are states reachable through outgoing edges, while its parents are the sources of incoming edges. States lacking outgoing edges are termed terminal states or sinks and are denoted as $x \in X$. It is essential for GFlowNets that users define the MDP in a way that ensures this graph remains acyclic, contains precisely one state without incoming edges, known as the initial state (source), and adheres to $R : X \to \mathbb{R} \geq 0$.

A complete trajectory is a sequence of states $\tau(s_0 \to s_1 \to \cdots \to s_n)$ starting from the source s_0 and leading to a sink s_n, with $(s_t \to s_{t+1}) \in A_{s_t}$ for all t. We use T to denote the set of all

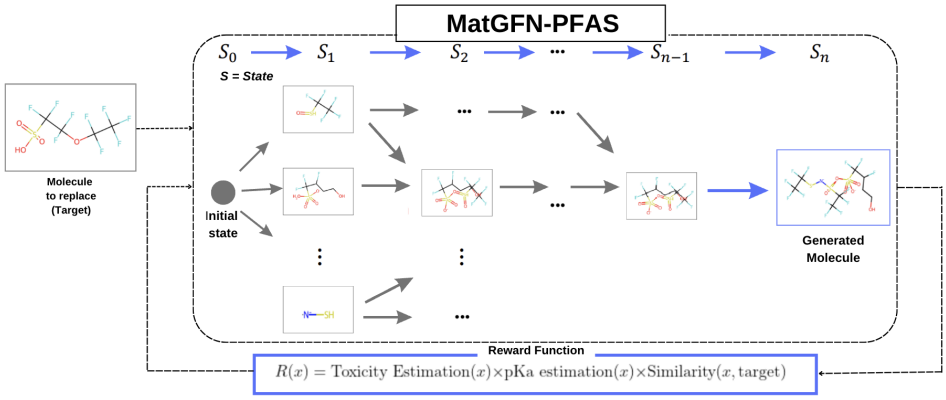

Figure 1: General architecture of the proposed MatGFN-PFAS approach.

complete trajectories. A trajectory flow is a non-negative function $F : T \to R \geq 0$, describing the unnormalized probability flowing along each complete trajectory τ from the source to a sink. For any state s, the state flow $F(s) = \sum_{\{\tau \in T: s \in T\}} F(\tau)$ quantifies the total unnormalized probability passing through state s. Additionally, for any edge $s \to s'$, a trajectory flow $F(\tau)$ is considered Markovian if there exist distributions $P_F(\cdot|s)$ over the children of every non-terminal state s, along with a constant Z, such that for any complete trajectory τ, we have $P_F(\tau) = F(\tau)/Z$, and $P_F(\tau = (s_0 \to s_1 \to \cdots \to s_n)) = \prod_{t=1}^{n} = P_F(s_t|s_{t-1})$. This $P_F(s_t|s_{t-1})$ is termed a forward policy, which can be used to sample complete trajectories from F.

GFlowNets are trained using stochastic gradient descent to optimize the learning objective on states or trajectories sampled from a training policy [27]. This training policy is typically a combination of P_F^θ and a uniform action policy, serving to encourage exploration during training. In the realm of Reinforcement Learning, GFlowNet training is considered off-policy. Importantly, GFlowNet training is a bootstrapping process where the current policy is utilized to sample new x at each training round. Since $R(x)$ is defined by the user, it is computed for each new x, and this set $\{x, R(x)\}$ is employed to update the GFlowNet policy.

In our specific case, $R(x)$ is defined as $R(x) =$ Toxicity prediction$(x) \times$ pKa prediction$(x) \times$ Tversky Similarity$(x, Target)$, with Target representing the PFAS molecule for which we seek a substitute. We use absolute values of pKa, and generated positive pKas are penalized, receiving the value 0.00001. The significance of a negative pKa value lies in ensuring the production of superacids, which find essential applications across various industries such as chip manufacturing [28]. While it may be challenging to eliminate their use entirely [28], the objective of the proposed approach is to generate safer alternatives.

2.2 MoLFormer

MoLFormer [24], is a large-scale masked chemical language model that processes inputs through a series of blocks that alternate between self-attention and feed-forward connections. MoLFormer was trained in a self-supervision manner with 1.1 billion molecules from PubChem and ZINC datasets and uses tokenization process, as detailed in [29]. The MoLFormer vocabulary includes 2362 unique chemical tokens. These tokens are used to fine-tune or retrain the MolFormer model. To reduce computation time, the sequence length has been limited to a range of 202 tokens as 99.4% percent of all 1.1 billion molecules contain less than 202 tokens.

MOLFORMER is equipped with a self-attention mechanism that allows the network to construct complex representations that incorporate context from across the sequence of SMILES. By transforming the sequence features into queries (q), keys (k), and value (v) representations, attention mechanisms can weigh the importance of different elements within the sequence. This enables the model to learn highly informative representations of the input data, making it a powerful tool for predicting

molecular properties. In this work, we used a version of the MoLFormer that was trained to predict toxicity, LD50 and pKa of PFAS elements [25]. The model reported 75% of accuracy on the toxicity task, and it is state-of-the-art in this domain.

2.3 MatGFN-PFAS training details

The following parameters have been used to train the MatGFN-PFAS for each of the 6 PFAS molecules studied in this experiment:

Table 1: MatGFN-PFAS agent parameters

Parameter	Value
Learning rate	5e-3
Epochs	35000
Mini batch size	5

Fig. 2 shows the loss trajectory while training the MatGFN-PFAS to generate molecules based on the studied SMILES **CC1CC(CC(F)(F)C(F)(F)OC(F)(F)C(F)(F)S(=O)(=O)O)OC1=O**, Fig. 2 also shows the logZ for the MatGFN-PFAS training. The other studied SMILES demonstrated similar behavior.

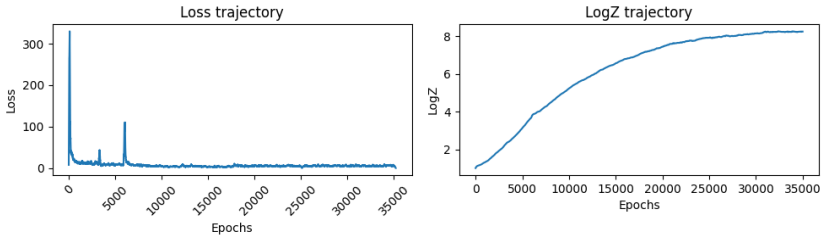

Figure 2: The figure illustrates the loss trajectory for the MatGFN-PFAS training. Losses are stabilized after epoch 6,000. LogZ for the MatGFN-PFAS training is also illustrated.

3 Results

To evaluate our proposed approach we selected a set of 6 PFAS SMILES which the LD50 measurement between 50-500 mg/kg, moderate toxicity according to the U.S. Environmental Protection Agency (EPA) [30]. The studied PFAS molecules have the structure defined as $R - CF_2OCF_2 - R'$, where R and R' can either be F, O, or saturated carbons. This is a PFAS definition stated by the U.S EPA agency [31]. Table 2 provides details about the studied PFAS molecules.

Table 2: Studied PFAS molecules which have the structure as $R - CF_2OCF_2 - R'$

SMILES	LD50 (mg/kg body weight)	pKA
CC1CC(CC(F)(F)C(F)(F)OC(F)(F)C(F)(F)S(=O)(=O)O)OC1=O	416.0	-3.12
O=C1CCC(CC(F)(F)C(F)(F)OC(F)(F)C(F)(F)S(=O)(=O)O)O1	406.0	-3.12
O=S(=O)(O)C(F)(F)C(F)(F)OC(F)(F)C(F)(F)C(F)(F)C(F)(F)OC(F)(F)C(F)(F)S(=O)(=O)O	272.0	-2.87
O=S(=O)(O)C(F)(F)C(F)(F)OC(F)(F)C(F)(F)C(F)(F)F	236.0	-3.22
O=S(=O)(O)C(F)(F)C(F)(F)OC(F)(F)C(F)(F)C1CC2CCC1C2	423.0	-3.12
O=S(=O)(O)C(F)(F)C(F)(F)OC(F)(F)C(F)(F)F	416.0	-3.25

Figure 3, illustrates the top 5 results in terms of Tversky similarity to the query molecule: **CC1CC(CC(F)(F)C(F)(F)OC(F)(F)C(F)(F)S(=O)(=O)O)OC1=O**. It is possible to note through the preliminary results that the proposed algorithm tries to generate molecules with low toxicity and pKa at the same time that it tries to maximize the similarity between the generated molecule and the target molecule.

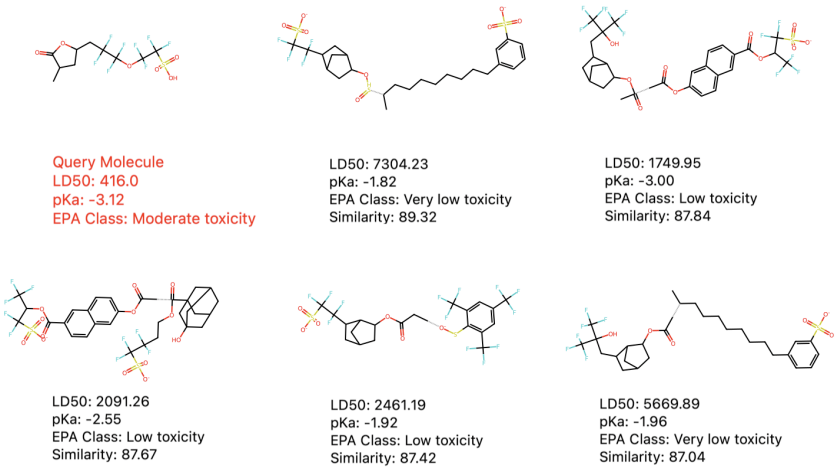

Figure 3: Top 5 generated molecules in terms of similarity score to the query SMILES: **CC1CC(CC(F)(F)C(F)(F)OC(F)(F)C(F)(F)S(=O)(=O)O)OC1=O**.

Initial results demonstrate that our algorithm exhibits a dual focus. Firstly, it strives to produce molecules with reduced toxicity and pKa levels, aligning with the goal of enhancing drug safety and minimizing potential side effects. Secondly, it places a strong emphasis on maximizing the similarity between the generated molecule and the target molecule. This dual objective highlights the algorithm's versatility and its potential to contribute significantly to drug discovery and design processes. By achieving a delicate balance between toxicity reduction and structural similarity. For the other studied molecules MatGFN-PFAS also generated toxic PFAS replacement candidates with low toxicity, low pKa, and similarity scores ranging from 85% to 95%. Detailed results for the other studied PFAS SMILES are included in the Supplementary Materials. The resulting datasets of this ongoing work are available at `https://ibm.box.com/v/MatGFN-PFAS-generated-datasets`.

The reported findings constitute an ongoing research effort, necessitating future investigations to validate the proposed framework's robustness. Emphasis is placed on testing the framework across an expanded and more complex PFAS dataset, encompassing diverse compounds to assess adaptability. Additionally, there is a call for the development of rigorous mechanisms to evaluate the quality of generated molecules, incorporating advanced criteria aligned with PFAS toxicity, structural fidelity, and environmental impact. These proposed avenues aim to refine the framework, positioning it as an effective tool for the generation of environmentally sustainable alternatives to toxic PFAS compounds.

4 Conclusion

In this paper, we introduce MatGFN-PFAS, an AI system leveraging GFlowNet and chemical foundation model, MoLFormer, for the purposeful generation of a diverse set of molecular candidates earmarked for the substitution of toxic per- and polyfluoroalkyl substances (PFAS). The architectural foundation of MatGFN-PFAS integrates a nuanced reward function, incorporating predictive indices such as LD50 and pKa, alongside the Tversky similarity metric. Our ongoing investigation substantiates the efficacy of this method, elucidating its proficiency in yielding molecules characterized by minimized toxicity and superacidic pKa values. Notably, MatGFN-PFAS is designed to concurrently optimize both structural fidelity and chemical congruence to the specified PFAS target.

For future research, our paramount objective resides in test the proposed framework on a expanded and more complex PFAS dataset, encompassing diverse compounds to assess adaptability. Furthermore, our commitment extends towards the refinement of the quality metrics pertaining to the generated molecular candidates. Subsequent endeavors are oriented towards the calibration of MatGFN-PFAS, ensuring its capacity not only to yield molecular candidates with commendable toxicity profiles and structural fidelity but also to align with the exacting standards requisite for robust PFAS replacement strategies. The resulting datasets generated by this ongoing investigation can be found at `https://ibm.box.com/v/MatGFN-PFAS-generated-datasets`.

References

[1] M. G. Evich, M. J. Davis, J. P. McCord, B. Acrey, J. A. Awkerman, D. R. Knappe, A. B. Lindstrom, T. F. Speth, C. Tebes-Stevens, M. J. Strynar *et al.*, "Per-and polyfluoroalkyl substances in the environment," *Science*, vol. 375, no. 6580, p. eabg9065, 2022.

[2] Z. Wang, A. M. Buser, I. T. Cousins, S. Demattio, W. Drost, O. Johansson, K. Ohno, G. Patlewicz, A. M. Richard, G. W. Walker *et al.*, "A new oecd definition for per-and polyfluoroalkyl substances," *Environmental science & technology*, vol. 55, no. 23, pp. 15 575–15 578, 2021.

[3] J. Glüge, M. Scheringer, I. T. Cousins, J. C. DeWitt, G. Goldenman, D. Herzke, R. Lohmann, C. A. Ng, X. Trier, and Z. Wang, "An overview of the uses of per-and polyfluoroalkyl substances (pfas)," *Environmental Science: Processes & Impacts*, vol. 22, no. 12, pp. 2345–2373, 2020.

[4] Z. Zhang, D. Sarkar, J. K. Biswas, and R. Datta, "Biodegradation of per-and polyfluoroalkyl substances (pfas): A review," *Bioresource technology*, vol. 344, p. 126223, 2022.

[5] S. F. Nakayama, M. Yoshikane, Y. Onoda, Y. Nishihama, M. Iwai-Shimada, M. Takagi, Y. Kobayashi, and T. Isobe, "Worldwide trends in tracing poly-and perfluoroalkyl substances (pfas) in the environment," *TrAC Trends in Analytical Chemistry*, vol. 121, p. 115410, 2019.

[6] A. O. De Silva, J. M. Armitage, T. A. Bruton, C. Dassuncao, W. Heiger-Bernays, X. C. Hu, A. Kärrman, B. Kelly, C. Ng, A. Robuck *et al.*, "Pfas exposure pathways for humans and wildlife: a synthesis of current knowledge and key gaps in understanding," *Environmental toxicology and chemistry*, vol. 40, no. 3, pp. 631–657, 2021.

[7] E. Panieri, K. Baralic, D. Djukic-Cosic, A. Buha Djordjevic, and L. Saso, "Pfas molecules: a major concern for the human health and the environment," *Toxics*, vol. 10, no. 2, p. 44, 2022.

[8] A. Ramírez Carnero, A. Lestido-Cardama, P. Vazquez Loureiro, L. Barbosa-Pereira, A. Rodríguez Bernaldo de Quirós, and R. Sendón, "Presence of perfluoroalkyl and polyfluoroalkyl substances (pfas) in food contact materials (fcm) and its migration to food," *Foods*, vol. 10, no. 7, p. 1443, 2021.

[9] J. L. Domingo and M. Nadal, "Human exposure to per-and polyfluoroalkyl substances (pfas) through drinking water: A review of the recent scientific literature," *Environmental research*, vol. 177, p. 108648, 2019.

[10] A. M. Calafat, L.-Y. Wong, Z. Kuklenyik, J. A. Reidy, and L. L. Needham, "Polyfluoroalkyl chemicals in the us population: data from the national health and nutrition examination survey (nhanes) 2003–2004 and comparisons with nhanes 1999–2000," *Environmental health perspectives*, vol. 115, no. 11, pp. 1596–1602, 2007.

[11] K. Roth, Z. Yang, M. Agarwal, W. Liu, Z. Peng, Z. Long, J. Birbeck, J. Westrick, W. Liu, and M. C. Petriello, "Exposure to a mixture of legacy, alternative, and replacement per-and polyfluoroalkyl substances (pfas) results in sex-dependent modulation of cholesterol metabolism and liver injury," *Environment International*, vol. 157, p. 106843, 2021.

[12] A. J. Blomberg, Y.-H. Shih, C. Messerlian, L. H. Jørgensen, P. Weihe, and P. Grandjean, "Early-life associations between per-and polyfluoroalkyl substances and serum lipids in a longitudinal birth cohort," *Environmental research*, vol. 200, p. 111400, 2021.

[13] S. Fragki, H. Dirven, T. Fletcher, B. Grasl-Kraupp, K. Bjerve Gützkow, R. Hoogenboom, S. Kersten, B. Lindeman, J. Louisse, A. Peijnenburg *et al.*, "Systemic pfos and pfoa exposure and disturbed lipid homeostasis in humans: what do we know and what not?" *Critical reviews in toxicology*, vol. 51, no. 2, pp. 141–164, 2021.

[14] R. Jin, R. McConnell, C. Catherine, S. Xu, D. I. Walker, N. Stratakis, D. P. Jones, G. W. Miller, C. Peng, D. V. Conti *et al.*, "Perfluoroalkyl substances and severity of nonalcoholic fatty liver in children: an untargeted metabolomics approach," *Environment international*, vol. 134, p. 105220, 2020.

[15] R. C. Lewis, L. E. Johns, and J. D. Meeker, "Serum biomarkers of exposure to perfluoroalkyl substances in relation to serum testosterone and measures of thyroid function among adults and adolescents from nhanes 2011–2012," *International journal of environmental research and public health*, vol. 12, no. 6, pp. 6098–6114, 2015.

[16] L. A. Darrow, A. C. Groth, A. Winquist, H.-M. Shin, S. M. Bartell, and K. Steenland, "Modeled perfluorooctanoic acid (pfoa) exposure and liver function in a mid-ohio valley community," *Environmental health perspectives*, vol. 124, no. 8, pp. 1227–1233, 2016.

[17] V. Gallo, G. Leonardi, B. Genser, M.-J. Lopez-Espinosa, S. J. Frisbee, L. Karlsson, A. M. Ducatman, and T. Fletcher, "Serum perfluorooctanoate (pfoa) and perfluorooctane sulfonate (pfos) concentrations and liver function biomarkers in a population with elevated pfoa exposure," *Environmental health perspectives*, vol. 120, no. 5, pp. 655–660, 2012.

[18] A. M. Mora, A. F. Fleisch, S. L. Rifas-Shiman, J. A. W. Baidal, L. Pardo, T. F. Webster, A. M. Calafat, X. Ye, E. Oken, and S. K. Sagiv, "Early life exposure to per-and polyfluoroalkyl substances and mid-childhood lipid and alanine aminotransferase levels," *Environment international*, vol. 111, pp. 1–13, 2018.

[19] Q. Qi, S. Niture, S. Gadi, E. Arthur, J. Moore, K. E. Levine, and D. Kumar, "Per-and polyfluoroalkyl substances activate upr pathway, induce steatosis and fibrosis in liver cells," *Environmental Toxicology*, vol. 38, no. 1, pp. 225–242, 2023.

[20] M. E. Andersen, B. Hagenbuch, U. Apte, J. C. Corton, T. Fletcher, C. Lau, W. L. Roth, B. Staels, G. L. Vega, H. J. Clewell 3rd *et al.*, "Why is elevation of serum cholesterol associated with exposure to perfluoroalkyl substances (pfas) in humans? a workshop report on potential mechanisms," *Toxicology*, vol. 459, p. 152845, 2021.

[21] T. Schillemans, I. Bergdahl, K. Hanhineva, L. Shi, C. Donat-Vargas, J. Koponen, H. Kiviranta, R. Landberg, A. Åkesson, and C. Brunius, "Associations of pfas-related plasma metabolites with cholesterol and triglyceride concentrations," *Environmental research*, vol. 216, p. 114570, 2023.

[22] M. Ateia, J. V. Buren, W. Barrett, T. Martin, and G. G. Back, "Sunrise of pfas replacements: A perspective on fluorine-free foams," *ACS Sustainable Chemistry & Engineering*, 2023.

[23] Y. Bengio, S. Lahlou, T. Deleu, E. J. Hu, M. Tiwari, and E. Bengio, "Gflownet foundations," *arXiv preprint arXiv:2111.09266*, 2021.

[24] J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, and P. Das, "Large-scale chemical language representations capture molecular structure and properties," *Nature Machine Intelligence*, vol. 4, no. 12, pp. 1256–1264, 2022.

[25] E. Soares, E. V. Brazil, K. F. A. Gutierrez, R. Cerqueira, D. Sanders, K. Schmidt, and D. Zubarev, "Beyond chemical language: A multimodal approach to enhance molecular property prediction," *arXiv preprint arXiv:2306.14919*, 2023.

[26] R. Kunimoto, M. Vogt, and J. Bajorath, "Maximum common substructure-based tversky index: an asymmetric hybrid similarity measure," *Journal of computer-aided molecular design*, vol. 30, pp. 523–531, 2016.

[27] M. W. Shen, E. Bengio, E. Hajiramezanali, A. Loukas, K. Cho, and T. Biancalani, "Towards understanding and improving gflownet training," *arXiv preprint arXiv:2305.07170*, 2023.

[28] C. K. Ober, F. Käfer, and J. Deng, "Review of essential use of fluorochemicals in lithographic patterning and semiconductor processing," *Journal of Micro/Nanopatterning, Materials, and Metrology*, vol. 21, no. 1, pp. 010 901–010 901, 2022.

[29] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, and A. A. Lee, "Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction," *ACS central science*, vol. 5, no. 9, pp. 1572–1583, 2019.

[30] K. Morris-Schaffer and M. J. McCoy, "A review of the ld50 and its current role in hazard communication," *ACS Chemical Health & Safety*, vol. 28, no. 1, pp. 25–33, 2020.

[31] E. P. A. EPA, "Procedures for chemical risk evaluation under the toxic substances control act (tsca) frl-7906-03-ocspp."

Supplementary Materials

Additional results

For each PFAS molecule studied in this work, we generated 1000 new replacement candidates. Below we present the top 5 generated molecules, ranked by their similarity scores to the PFAS molecules utilized as queries for MatGFN-PFAS:

1. Fig.4 illustrates the top generated molecules in terms of similarity to the query: **O=C1CCC(CC(F)(F)C(F)(F)OC(F)(F)C(F)(F)S(=O)(=O)O)O1**, $LD50 = 406.0$, $pKa = -3.12$

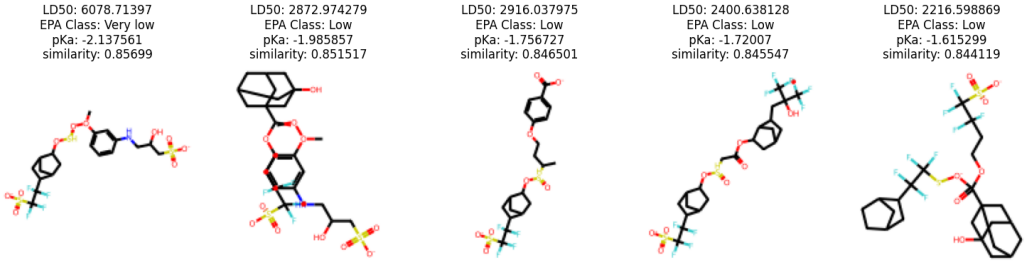

Figure 4: Top 5 results in terms of similarity to the query SMILE: **O=C1CCC(CC(F)(F)C(F)(F)OC(F)(F)C(F)(F)S(=O)(=O)O)O1**.

2. Results for the query: **O=S(=O)(O)C(F)(F)C(F)(F)OC(F)(F)C(F)(F)C(F)(F)C(F)(F)OC(F)(F)C(F)(F)S(=O)(=O)O**, $LD50 = 272.0$, $pKa = -2.87$, are illustrated by Fig.5.

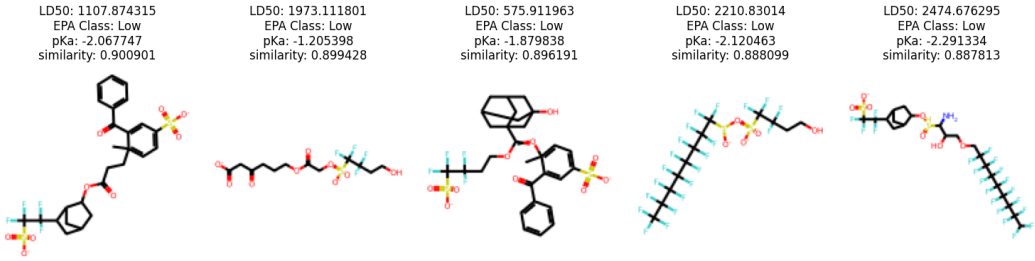

Figure 5: Top 5 results in terms of similarity to the query SMILE: **O=S(=O)(O)C(F)(F)C(F)(F)OC(F)(F)C(F)(F)C(F)(F)C(F)(F)OC(F)(F)C(F)(F)S(=O)(=O)O**.

3. The top results in terms of similarity to the query: **O=S(=O)(O)C(F)(F)C(F)(F)OC(F)(F)C(F)(F)C(F)(F)F**, $LD50 = 236.0$, $pKa = -3.22$ are illustrated in Fig.6.

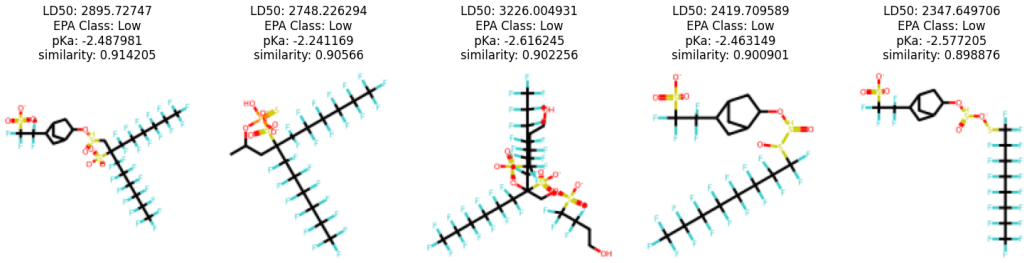

Figure 6: Top 5 results in terms of similarity to the query SMILE: **O=S(=O)(O)C(F)(F)C(F)(F)OC(F)(F)C(F)(F)C(F)(F)F**.

4. For the query **O=S(=O)(O)C(F)(F)C(F)(F)OC(F)(F)C(F)(F)C1CC2CCC1C2**, $LD50 = 423.0$, $pKa = -3.12$, the top 5 generated results in terms of similarity score are demonstrated by Fig.7.

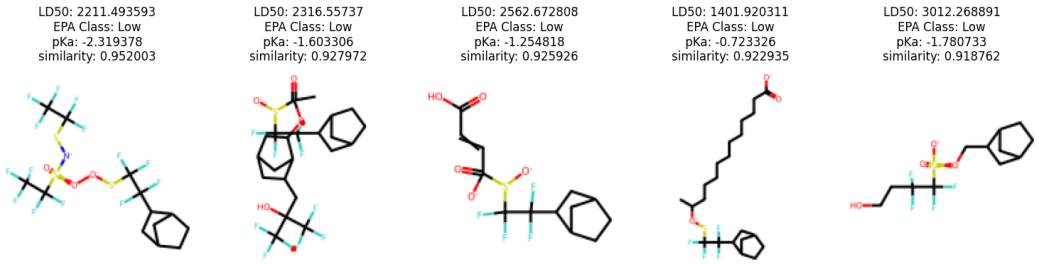

Figure 7: Top 5 results in terms of similarity to the query SMILE:
O=S(=O)(O)C(F)(F)C(F)(F)OC(F)(F)C(F)(F)C1CC2CCC1C2.

5. Finally, for the query **O=S(=O)(O)C(F)(F)C(F)(F)OC(F)(F)C(F)(F)F**, $LD50 = 416.0$, $pKa = -3.25$, the results are illustrated by Fig.8.

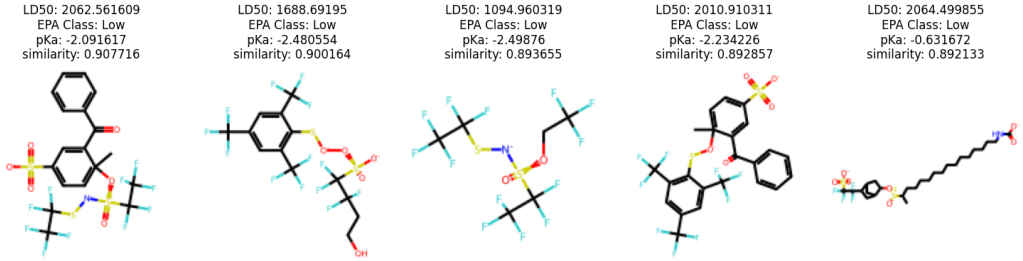

Figure 8: Top 5 results in terms of similarity to the query SMILE:
O=S(=O)(O)C(F)(F)C(F)(F)OC(F)(F)C(F)(F)F.

Through these results we can observe that the proposed MatGFN-PFAS framework consistently tries to generate replacement molecules that follows the constraints of having low toxicity in terms of LD50 measurements, negative pKa values (superacids), and high structural similarities to the query molecules, keeping the main the properties of the studied molecules.

These initial findings of this ongoing study demonstrate the effectiveness of the proposed MatGFN-PFAS approach in generating possible toxic PFAS replacements. Future work should concentrate on improving the results as well as evaluating the quality of the generated molecules. The resulting datasets are available at `https://ibm.box.com/v/MatGFN-PFAS-generated-datasets`.