CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Reinforcement Learning-Guided Generative Design of Surfactant Molecules

Exploring Reward Functions for Multi-Objective Molecular Optimisation

Master's thesis in Complex adaptive systems

Hannes Öhman

Department of Computer Science and Engineering

CHALMERS UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Gothenburg, Sweden 2026

Master's thesis 2026

Reinforcement Learning-Guided Generative Design of Surfactant Molecules

Exploring Reward Functions for Multi-Objective Molecular Optimisation

Hannes Öhman

UNIVERSITY OF GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2026

Reinforcement Learning-Guided Generative Design of Surfactant Molecules
Exploring Reward Functions for Multi-Objective Molecular Optimisation
Hannes Öhman

Supervisor: Richard Beckmann, Data Science och AI
Examiner: Rocío Mercado Oropeza, Data Science och AI

Master's Thesis 2026
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2026

Reinforcement Learning-Guided Generative Design of Surfactant Molecules
Exploring Reward Functions for Multi-Objective Molecular Optimisation

Hannes Öhman
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

Surfactants are chemical compounds with amphiphilic molecular structures that find application across a wide range of industries, from cleaning products and food processing to pharmaceuticals and waterproof textiles. Despite their ubiquity, many surfactants in current industrial use raise significant environmental concerns, and the physicochemical performance of even commonly used compounds often falls far short of what is theoretically achievable. Designing better surfactants through conventional experimental methods is slow and costly, motivating the development of computational approaches capable of navigating the vast space of possible molecular structures more intelligently.

This thesis presents a computational framework for the inverse design of surfactant molecules with improved physicochemical properties, combining deep generative modelling with reinforcement learning and active learning. The framework is built on the REINVENT4 platform, employing a recurrent neural network pretrained on a large chemical database and subsequently fine-tuned on the SurfPro surfactant dataset through transfer learning. Reinforcement learning is then applied to guide generation toward molecules with favourable values of critical micelle concentration (pCMC) and surface tension, using surrogate models as the property oracle. Three distinct optimisation strategies are investigated and compared: iterative retraining, Pareto Gradient scoring, and Origin Pull scoring.

The results demonstrate that the framework reliably generates chemically valid and novel molecules that extend beyond the Pareto front of the SurfPro training dataset according to surrogate model predictions. A key finding is that balanced multi-objective scoring functions both outperform single-objective weightings and better preserve chemical diversity throughout training. However, a recurring failure mode termed here the hydrocarbon trap is identified, in which the optimisation converges on structurally degenerate

molecules that lack the amphiphilic head-tail architecture required of functional surfactants. A post-processing filter based on physicochemical descriptors derived from the SurfPro dataset is introduced to screen out these spurious candidates and recover the viable subset of generated molecules.

Taken together, the results constitute a proof of concept that reinforcement learning-guided generative models are capable of exploring surfactant chemical space in a principled and targeted manner. While the generated candidates require experimental validation before their practical utility can be confirmed, this framework represents a promising foundation for future, more refined computational campaigns aimed at discovering next-generation surfactants.

Keywords: Industrial Chemistry, Data Science, Generative Models, AI, Computer Science, Surfactants, Molecular Engineering, Molecular Design

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Richard Beckmann, for his guidance, patience, and support throughout this project. His insight and constructive feedback shaped this work at every stage, and the direction this thesis ultimately took owes a great deal to his involvement.

I would also like to thank my examiner, Rocío Mercado Oropeza, for the time and care invested in reviewing this work and for providing thoughtful and valuable feedback.

Finally, and most importantly, I would like to thank my friends and family. Their support throughout my years of study has meant more than I can adequately express here.

Hannes Öhman, Gothenburg, 2026-06-09

Contents

List of Acronyms

AI	Artificial Intelligence.
CMC	Critical Micelle Concentration.
HV	Hypervolume.
LogP / MolLogP	(Molecular) Logarithm of the Partition Coefficient.
LSTM	Long Short-Term Memory.
pCMC	(negative logarithm of) Critical Micelle Concentration.
RL	Reinforcement Learning.
RNN	Recurrent Neural Network.
SMILES	Simplified Molecular Input Line Entry System.
TPSA	Topological Polar Surface Area.

List of Figures

List of Tables

1

Introduction

1.1 What Surfactants Are and Why They Matter

Surfactants are a class of chemical compounds characterised by their amphiphilic molecular structure, one end of the molecule is hydrophilic and attracted to water, while the other is hydrophobic and repels it. This dual nature allows surfactants to act at the interface between two otherwise immiscible phases, such as oil and water, making them indispensable across a wide range of industries. Applications include detergents and cleaning products, food processing and packaging, cosmetics, pharmaceuticals, and waterproof textiles, among many others[1], [2], [3]. The global ubiquity of surfactants in both industrial and consumer products makes the properties of these molecules a matter of significant practical importance.

1.2 The Need for Better and Safer Surfactants

Despite their widespread utility, many surfactants in common use today raise significant environmental and human health concerns. Across the broader landscape of synthetic surfactants, issues of environmental persistence, aquatic toxicity, and poor biodegradability are common[4], [5]. Even among surfactants not subject to regulatory restrictions, there is a wide spectrum of performance: many compounds used in practice are far from optimal with respect to properties such as critical micelle concentration and surface tension reduction, the very properties that determine how effectively a surfactant can fulfill its task.

This creates a clear and broad motivation: the chemical industry needs access to surfac-

tants that are not only more environmentally benign and biodegradable, but also more effective. The challenge is therefore not merely to find drop-in replacements for any one harmful class, but to push the overall frontier of surfactant design towards molecules with superior functional properties and a reduced environmental footprint. Achieving this in a principled way, rather than by slow experimental trial and error, requires new computational approaches capable of efficiently navigating the vast space of possible molecular structures.

1.3 Why Designing Novel Surfactants Is Hard

The central challenge in discovering better surfactant molecules lies in the sheer scale of chemical space. The number of possible drug-like or material-like molecules that can be constructed from common atoms is estimated to exceed 10^{60} [6], a number so large that exhaustive exploration is not merely impractical but fundamentally impossible. Traditionally, the discovery of new surfactant molecules has relied on laboratory synthesis followed by experimental characterisation. A process that is both time-consuming and expensive, and which offers no guarantee of finding candidates with the desired properties. Even when guided by chemical intuition, this trial-and-error approach can only probe a vanishingly small fraction of the available chemical space. The result is that the known universe of characterised surfactants remains extremely limited relative to what likely exists, and the probability of stumbling upon an optimal replacement by conventional means is low.

1.4 Machine Learning as a Solution

Recent advances in machine learning have opened up a fundamentally different approach to molecular design[7], [8]. Rather than synthesising and testing candidates one by one, machine learning models can be trained to learn the underlying patterns that distinguish useful molecules from poor ones, and then leverage that knowledge to propose new candidates computationally. Generative models are particularly well suited to this task. Unlike virtual screening approaches, which can only evaluate molecules that already exist in a predefined library, generative models are capable of constructing entirely new molecular structures, effectively learning a map of chemical space and navigating it intelligently. When combined with reinforcement learning, such models can be guided towards regions

of chemical space associated with desirable properties, dramatically accelerating the search process. Despite the growing application of these methods in drug discovery and materials science[9], [10], [11], their use in surfactant design remains largely unexplored[12], leaving a clear opportunity for the kind of investigation undertaken in this work.

1.5 Scope, Target Properties, and Contribution

This thesis presents a computational framework for the inverse design of surfactant molecules with optimised physicochemical properties, combining deep generative modelling with reinforcement learning and active learning. Two properties are chosen as the primary optimisation targets: the critical micelle concentration, expressed as its negative logarithm *pCMC*, and surface tension reduction. Both are fundamental to the practical utility of a surfactant. As for the attributes, pCMC quantifies the efficiency with which a molecule forms micelles, with higher values indicating that micelle formation occurs at lower concentrations, while surface tension reduction captures the ability of the surfactant to lower the surface tension of a liquid, which is central to most of its functional applications. These two properties were selected not only for their chemical relevance but also because accurate surrogate models capable of predicting them directly from molecular SMILES strings are available, making them well suited for integration into a reinforcement learning workflow. The contribution of this work is best understood as a proof of concept: a demonstration that generative models, guided by reinforcement learning and active learning, are capable of proposing theoretically viable surfactant candidates that improve upon the known property trade-offs in the SurfPro dataset and that may, subject to experimental validation, represent a step towards a new generation of higher-performing, environmentally conscious surfactants. The thesis does not address the synthesisability of the proposed molecules, nor does it identify specific compounds as definitive replacement candidates, these questions are left for future experimental work.

1.6 Research Questions

The following research questions guide this investigation.

- How can a generative model be guided to good, industrially viable surfactant molecules?

- What kind of reward function is best suited for this notoriously difficult optimization problem?
- Can we find a surfactant that outperforms state of the art surfactants in terms of pCMC and surface tension?

Together, these questions frame both the technical and practical ambitions of the work, and the results and discussion are structured around providing answers to each.

2

Theory

2.1 Molecular Representation Using SMILES

For a machine learning model to reason about molecules, those molecules must first be expressed in a format that a computer can process. One of the most widely used representations for this purpose is the Simplified Molecular Input Line Entry System (SMILES) [13], which encodes the full chemical structure of a molecule as a linear string of characters.

In a SMILES string, atoms are represented by their chemical symbols, carbon as C, nitrogen as N, oxygen as O, and so on, and bonds between atoms are encoded by special characters, with single bonds implied by adjacency and double or aromatic bonds indicated explicitly. Branching in the molecular graph is represented using parentheses, and ring structures are encoded by assigning matching numerical indices to the two atoms that form the ring closure. As a simple example, ethanol (CH_3CH_2OH) is written as `CCO`, and benzene as `c1ccccc1`, where the lowercase letters indicate aromatic atoms and the indices 1 mark the ring closure.

The key property that makes SMILES particularly well suited for use with generative sequence models is that it reduces a complex two-dimensional molecular graph to a one-dimensional sequence of tokens. This means that the problem of generating a new molecule can be reframed as the problem of generating a valid string, a task that sequence models such as recurrent neural networks are specifically designed to handle. In practice, a vocabulary of tokens is defined covering all atom types, bond types, and structural characters that appear in the training data, and molecules are tokenised into sequences over this vocabulary before being passed to the model.

2.2 RNN-based Generative Models for Chemistry

Recurrent neural networks (RNNs) are a class of neural network designed to process sequential data by maintaining a hidden state that is updated at each step of the sequence [14]. At each step, the model receives the current input token and its previous hidden state, and produces both an updated hidden state and a probability distribution over the next token. This allows the model to condition each prediction on the full history of previously generated tokens, making RNNs naturally suited to tasks where context and order matter.

In the context of molecular generation, an RNN is trained to model the probability distribution over valid SMILES strings. During training, the model is presented with known SMILES sequences token by token and learns to assign high probability to chemically valid continuations at each step. After training, new molecules can be generated by sampling from the model autoregressively: a start token is provided, the model samples the next token from its predicted distribution, that token is fed back as input, and the process repeats until an end token is produced. The result is a complete SMILES string representing a generated molecule.

A key advantage of RNNs for SMILES generation is their ability to handle the long-range dependencies inherent in molecular strings. SMILES encodes structural features such as ring closures and branching points using paired characters that may be separated by many tokens in the sequence, for example, a ring opened early in the string must be closed at a matching position later. The hidden state of an RNN carries memory of all previously generated tokens, allowing the model to respect these dependencies and generate structurally consistent molecules. In practice, the Long Short-Term Memory (LSTM) variant of the RNN is commonly used for this task, as its gating mechanism allows it to retain relevant information over longer sequences more reliably than a standard RNN [14][15].

2.3 Policy-Gradient Reinforcement Learning for Sequence Generation

Reinforcement learning (RL) is a framework in which an agent learns to take actions in an environment by receiving reward signals that indicate how desirable the outcomes of those actions are [16]. The agent's goal is to learn a policy, a mapping from states to actions, that maximises the expected cumulative reward.

In the context of molecular generation with REINVENT, the RNN constitutes the policy. At each step of sequence generation, the model's current hidden state represents the state, the sampled token represents the action, and the policy is defined by the probability distribution the model assigns over the vocabulary of tokens. Once a complete SMILES string has been generated, it is evaluated by the scoring function and assigned a scalar reward. This reward is then used to update the model weights via a policy gradient method.

Policy gradient methods optimise the policy by estimating the gradient of the expected reward with respect to the model parameters. Intuitively, sequences that received high rewards are made more probable, and sequences that received low rewards are made less probable. REINVENT specifically employs an augmented likelihood formulation, in which the training signal combines the reward with a penalty for diverging from the prior. This keeps the agent policy from collapsing onto a narrow set of high-scoring but chemically implausible structures, a phenomenon known as mode collapse, and ensures that the generated molecules remain diverse and chemically meaningful throughout training.

2.4 Multi-Objective Reward Design and the Pareto Front

In many molecular design problems, optimising a single property is insufficient, multiple properties must be improved simultaneously, and these properties often trade off against one another. Designing a scoring function that captures this multi-objective nature requires care.

One approach to combining multiple objectives into a single scalar reward is the weighted geometric mean, defined as:

$$\text{score} = \left(\prod_{i=1}^{n} x_i^{w_i} \right)^{1/\sum_{i=1}^{n} w_i}$$

where x_i are the predicted scores of the individual scoring functions for each property, and w_i are non-negative weights. This formulation has the property that a very low score on any individual objective substantially depresses the overall score, encouraging balanced improvement across objectives rather than extreme performance on a single one.

Even with a well-designed scoring function, a single scalar reward cannot fully capture the trade-off structure of a multi-objective problem. The Pareto front provides a more complete picture. Given a set of molecules evaluated on two properties, a molecule A is said to dominate molecule B if A is at least as good as B on all objectives and strictly better on at least one. The Pareto front is the set of all non-dominated candidates, those for which no other candidate in the set is better on all objectives. Visually, when candidates are plotted in two-dimensional property space, the Pareto front forms the boundary of the point cloud closest to the optimal corner.

The quality of a generated Pareto front can be quantified using the dominated hypervolume indicator, which measures the volume of property space enclosed between the Pareto front and a fixed reference point. Improvements in this metric indicate that the generated set of candidates collectively represents a better coverage of the optimal trade-off region.

2.5 Active Learning

Active learning is a machine learning paradigm in which a model iteratively selects the most informative data points for labelling, rather than being trained on a fixed dataset [17]. The core idea is that by strategically choosing which examples to evaluate, guided by what the model has already learned, it is possible to achieve strong performance with significantly fewer labelled examples than passive learning on a randomly sampled dataset would require.

In the context of molecular design, active learning takes the form of a closed-loop process in which a generative model and an oracle, a function that evaluates molecular properties,

interact iteratively. At each iteration, the model generates a batch of candidate molecules, the oracle evaluates their properties, and the highest-value candidates are incorporated into the training set before the model is updated for the next iteration. This allows the model to progressively focus its exploration on the most promising regions of chemical space, improving sample efficiency by concentrating oracle calls where they are most likely to yield useful information.

A key distinction exists between active learning and standard reinforcement learning when applied to generative models. Reinforcement learning updates only the agent's policy weights, shifting its sampling behaviour within a fixed learned space. Active learning, by contrast, modifies the underlying model through repeated retraining on an expanding dataset, progressively embedding newly discovered high-performing structures more deeply into the model's learned distribution. This allows the model's conception of the search space itself to evolve across iterations, rather than merely its sampling preferences within a static space.

3

Methods

3.1 Overview

This work aims to develop a computational framework for the inverse design of surfactant molecules with optimised physicochemical properties. The approach combines deep generative modelling with reinforcement learning and active learning in order to efficiently explore chemical space and identify candidate molecules that improve upon known trade-offs between critical micelle concentration (pCMC) and surface tension.

The workflow is structured into four sequential stages. First, a pretrained generative model (the prior) is adapted to the domain of surfactants through transfer learning on the SurfPro dataset. Second, a weight sweep across 21 scoring function configurations is conducted to characterise the trade-off between pCMC and surface tension, and to identify a balanced baseline configuration. Third, the baseline model is compared against three distinct goal-directed optimisation strategies: *Iterative Retraining*, *Pareto Gradient*, and *Origin Pull*. Each representing a different approach to biasing generation towards improved regions of property space. Fourth, results are evaluated both against the SurfPro Pareto front and against an unguided sampling baseline, to establish that reinforcement learning provides a meaningful improvement over simple post-hoc selection from the transfer-learned model.

This pipeline enables both distribution learning, capturing the structure of known surfactants, and goal-directed optimisation, pushing beyond the training distribution towards improved performance

3.2 Model Architecture

The generative model used in this work is built on the REINVENT4 framework [18], an open-source platform for goal-directed molecular design. REINVENT4 supports RNN and transformer-based architectures embedded within transfer learning, reinforcement learning, and curriculum learning workflows. The RNN-based architecture was selected for this work.

REINVENT operates with two model instances: a fixed prior and an adjustable agent. The prior is a generative model pre-trained on a large and chemically diverse corpus, in this work that is the PubChem database, and captures general chemical syntax and structural patterns across a broad region of chemical space. The agent begins as a copy of the prior and is subsequently updated during reinforcement learning to bias generation towards desirable molecules. Crucially, the prior remains active throughout training as a regulariser: the agent's log-likelihood of generating a given molecule is constrained to remain close to that of the prior via an augmented likelihood formulation, preventing the agent from drifting too far into implausible or degenerate regions of chemical space. This regularisation is essential for maintaining chemical validity and diversity during optimisation.

The model learns a probability distribution over valid SMILES strings by sequentially predicting the next token in a sequence, thereby generating complete molecular structures autoregressively, as described in Section 2.2. The prior was selected for its compatibility with the atom types and structural motifs present in the SurfPro dataset. Training of the fine-tuned model was performed using a batch size of 256 and an initial learning rate of 1×10^{-4}, with exponential decay applied during training.

Property predictions for pCMC and surface tension were obtained using pre-trained surrogate models from Beckmann et al. [19]. These models take as input a set of approximately 50 molecular descriptors generated with RDKit [20], covering a broad range of structural and physicochemical features. The surrogate models were used both to fill in missing property values in the SurfPro dataset (see Section 4.1) and as the oracle during reinforcement learning, evaluating each generated molecule and returning a reward signal. Predicted property values were normalised using min-max scaling based on the range observed in

the SurfPro dataset, with the same scaling parameters applied consistently to generated molecules. Since both pCMC and surface tension are minimisation objectives, lower values indicate better surfactant performance, the normalised values were subsequently inverted by computing (1 – normalised value), such that a lower predicted property value maps to a higher score. This inversion ensures that the scoring function correctly rewards molecules with desirable low values of both properties.

3.3 Scoring Function Design

The scoring function used to guide molecule generation during reinforcement learning combines two objectives, pCMC and surface tension, using the weighted geometric mean formulation introduced in Section 2.4.

A known risk in surrogate-model-based reinforcement learning is reward hacking [21], wherein the generative model learns to exploit weaknesses or blind spots in the proxy reward rather than genuinely optimising the underlying physical objective. In the present setting, this risk is non-trivial: the surrogate models predict pCMC and surface tension from a fixed set of approximately 50 molecular descriptors that do not directly encode amphiphilicity. It is therefore possible, in principle, for a molecule to achieve favourable descriptor values, and thus a high predicted score, without possessing the hydrophilic head group and hydrophobic tail that are structurally necessary for genuine surfactant behaviour. The post-processing filter introduced in Section 3.6 is designed to screen out the most severely degenerate outputs after generation. However, the more fundamental mitigation would be to incorporate amphiphilicity-aware constraints directly into the scoring function, a direction left for future work. Readers should bear this limitation in mind when interpreting the optimisation results in Chapter 4.

To characterise the trade-off between the two objectives and identify a suitable baseline configuration, a weight sweep was conducted across 21 evenly spaced configurations of weight ratios between pCMC and surface tension, ranging from (0.00, 1.00) to (1.00, 0.00) in steps of 0.05. Results across this sweep were broadly similar, and a balanced weighting of (0.50, 0.50) was selected as the baseline configuration for subsequent experiments, reflecting equal importance assigned to both properties.

3.4 Training Procedure

3.4.1 Transfer Learning

The pretrained prior model was fine-tuned on the SurfPro dataset using supervised learning. This step adapts the general chemical knowledge of the prior to the specific structural and compositional patterns of surfactant molecules, producing a model whose unconditional output distribution is concentrated in the relevant region of chemical space.

3.4.2 Unguided Sampling Baseline

To establish a meaningful point of comparison for the reinforcement learning experiments, an unguided sampling baseline was constructed. A large batch of molecules was sampled from the transfer-learned model without any reinforcement learning or scoring function guidance. Each sampled molecule was then evaluated using the surrogate models, and the top-scoring molecules were retained. This baseline represents the best performance achievable through sampling alone, without active optimisation, and is used throughout the results to demonstrate the additional value provided by reinforcement learning.

3.4.3 Reinforcement Learning

Following transfer learning, the model was further optimised using reinforcement learning (RL) to bias generation towards high-scoring molecules. At each step, a batch of 256 molecules was generated and evaluated using the scoring function, with the resulting score used as a reward signal to update the model parameters via the policy gradient approach described in Section 2.3. A learning rate of 1×10^{-4} was used to provide stable optimisation. Each run consisted of 150 RL steps, corresponding to 38400 oracle calls in total. To improve the reliability of results, 10 parallel runs were conducted for each configuration, and reported metrics were aggregated across these runs. To ensure reproducibility, random seeds were fixed for each parallel run, with the seed for each run set equal to its index (i.e. seeds 0 through 9).

Three distinct reinforcement learning strategies were investigated, each representing a different approach to guiding generation towards improved regions of the pCMC-surface

tension property space.

3.4.3.1 Iterative Retraining

In this strategy, the 150 RL steps were divided into a number of generations, either 3, 5, or 10, with the steps distributed evenly across generations (e.g. 50 steps per generation for 3 generations). The iterative retraining strategy constitutes an active learning loop in which the model iteratively generates candidates, evaluates them, and incorporates the most optimal ones into the training set before the next round of optimisation. For our implementation, the most optimal molecules were defined as being the molecules on the Pareto-front of the generated molecules in the generation. These optimal molecules were then used to augment the SurfPro dataset. A new model was then created by transfer-learning on the original prior with this expanded dataset before the next generation of RL began. This closed-loop approach is intended to improve sample efficiency by allowing the model to refine its understanding of the property landscape and progressively focus exploration on the most promising regions of chemical space, rather than exploring uniformly across the full distribution.

3.4.3.2 Pareto Gradient

In this strategy, the Pareto front of all previously generated molecules was updated at every RL step. Each newly generated molecule was scored not only on its predicted property values but also on its position relative to the current Pareto front. Molecules that lie on or beyond the Pareto front, that is, molecules that dominate at least one previously generated molecule, received the maximum score. Molecules that did not reach the front received a score that decreased linearly with their distance from it. This continuous Pareto-based reward signal encourages the model to generate molecules that actively extend the frontier rather than simply improving average scores.

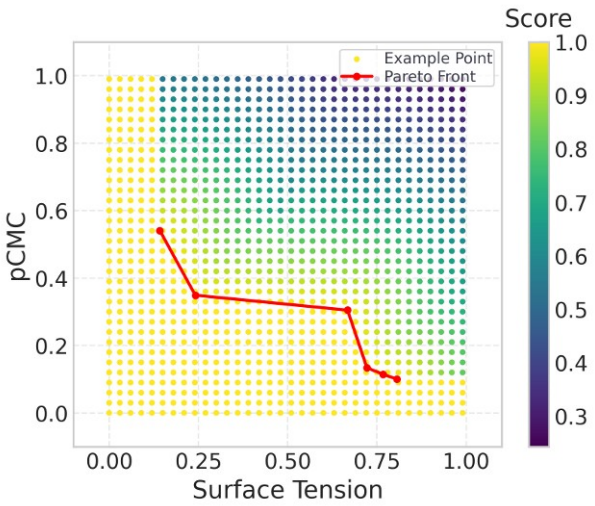

Figure 3.1: The scoring landscape of the *Pareto Gradient* scoring function in normalised pCMC - surface tension space. Each point represents a hypothetical molecule position in the two-dimensional property space, colour-coded by its assigned score. The red line marks the current Pareto front, molecules lying on or beyond the front receive the maximum score of 1.0, while molecules behind the front are scored in inverse proportion to their linear distance from it.

3.4.3.3 Origin Pull

This strategy is similar in structure to the Pareto Gradient but uses a different geometric reference point. Rather than scoring molecules based on their distance to the Pareto front, molecules are scored based on their distance to the origin of the property space. This encourages the model to generate molecules that are simultaneously extreme in both objectives, effectively pulling the distribution towards the corner of the property space where both pCMC and surface tension are minimised.

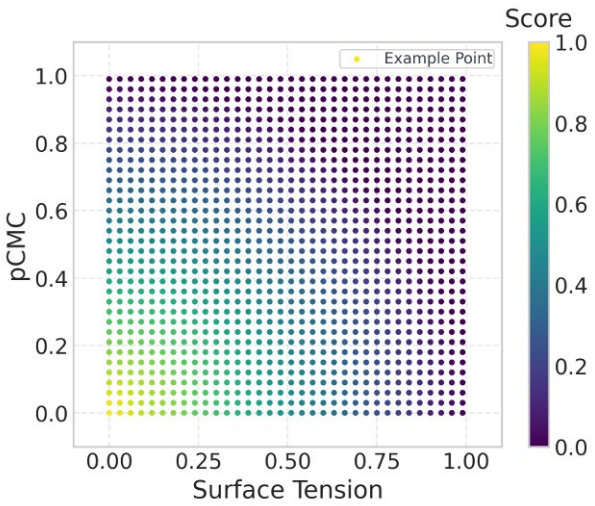

Figure 3.2: The scoring landscape of the *Origin Pull* scoring function in normalised pCMC - surface tension space. Each point represents a hypothetical molecule position in the two-dimensional property space, colour-coded by its assigned score. Scores decrease continuously with increasing Euclidean distance from the origin, rewarding molecules that are simultaneously low in both pCMC and surface tension.

3.5 Evaluation metrics

Model performance was evaluated using both generative quality metrics and optimisation-specific metrics.

For unconditional generation, the following metrics were used:

- Validity: fraction of chemically valid molecules
- Uniqueness: fraction of non-duplicate molecules
- Novelty: fraction of molecules not present in the training set

For goal-directed optimisation:

- **Dominated hypervolume:** the volume of property space dominated by the set of generated molecules relative to a fixed reference point, used to quantify the overall quality and spread of the Pareto front.
- **Distribution of generated molecules in property space:** visualised to assess the shift in the generated distribution relative to the SurfPro dataset.
- **Comparison against the unguided sampling baseline:** to establish the contribution of reinforcement learning over post-hoc selection alone.

- Improvement on the Pareto front relative to the SurfPro dataset
- Maximum achieved scores

3.6 Post Processing

To further narrow down the large quantity of generated molecules, we apply a strict set of filters based on structural and physico-chemical properties. The filter relies on two descriptors computed directly from molecular structure using RDKit. The first is a newly defined metric termed the *Branched Carbon Ratio*, computed as the number of tertiary and quaternary carbon atoms in a molecule divided by the total number of heavy atoms. Dividing by total heavy atom count normalises the metric for molecular size, yielding a ratio that is independent of molecular scale. The second descriptor is MolLogP, the WildmanCrippen partition coefficient, which estimates the lipophilicity of a molecule based on its atomic contributions.

Filter bounds for both descriptors were derived from the SurfPro dataset. The Branched Carbon Ratio and MolLogP were computed for all 1551 molecules in the preprocessed dataset, and the minimum and maximum observed values were taken as the lower and upper bounds respectively, defining the physicochemical envelope of known, experimentally characterised surfactants. A generated molecule is retained if and only if both its Branched Carbon Ratio and its MolLogP fall within the corresponding SurfPro bounds simultaneously. Any molecule violating either criterion is discarded. This conjunctive filter ensures that retained candidates are consistent with the reference data in both branching character and overall lipophilicity, providing a principled viability screen without imposing any structural template or substructure constraint on the generated molecules.

4

Results

4.1 Dataset

The SurfPro dataset [22] was used as the primary data source. It consists of 1623 surfactant molecules collected from experimental literature across multiple publications, covering a diverse range of surfactant classes including gemini cationic, non-ionic, anionic surfactants, among others. Each molecule is annotated with physicochemical properties including the critical micelle concentration (CMC).

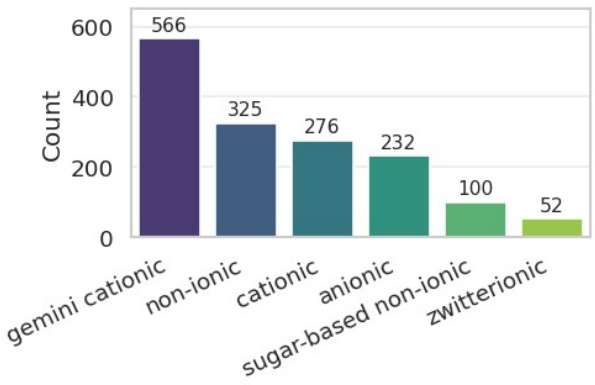

Figure 4.1: Bar chart showing the number of molecules belonging to each surfactant class in the SurfPro dataset.

As shown in Fig. 4.1, the dataset is dominated by gemini cationic surfactants (566 molecules, approximately 36% of the total), followed by non-ionic (325), cationic (276), anionic (232), sugar-based non-ionic (100), and zwitterionic (52) surfactants. This class imbalance reflects the broader literature from which the dataset was compiled, in which gemini surfactants have been extensively characterised owing to their enhanced surface activity relative to their single-chain counterparts. [22]

Not all molecules in the dataset included measurements for both target properties. Where surface tension or pCMC values were absent, they were computed using existing surrogate models [19]. This ensured that all molecules used in training and evaluation had both properties available, while remaining consistent with the surrogate-based evaluation used throughout the optimisation pipeline.

After preprocessing to remove invalid and duplicate entries, the dataset was reduced to 1551 molecules. This step ensures that the training data reflects chemically valid and non-redundant structures, which is essential for stable generative modelling.

All molecules were represented using canonical SMILES strings generated with RDKit [20], as described in Section 2.1

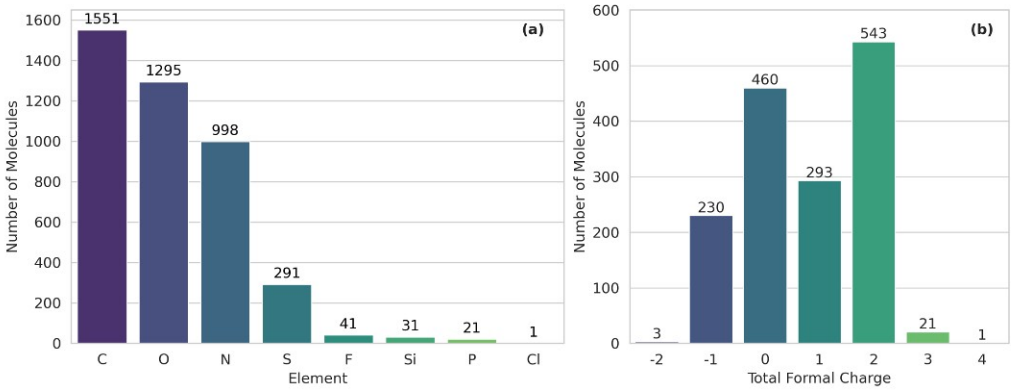

Figure 4.2: Elemental composition and formal charge distribution of the preprocessed SurfPro dataset. **(a)** Bar chart of the number of molecules containing each element. **(b)** Bar chart of the number of molecules carrying each value of total formal charge, ranging from -2 to +4.

Fig. 4.2 characterises the elemental composition and formal charge distribution of the preprocessed dataset. The elemental breakdown in panel **a** shows that carbon (present in all 1551 molecules), oxygen (1295), and nitrogen (998) are the most prevalent elements, consistent with the organic, amphiphilic nature of surfactant structures. Heteroatoms including sulfur (291), fluorine (41), silicon (31), and phosphorus (21) are represented in smaller numbers, reflecting the chemical breadth of the dataset across different surfactant classes. The formal charge distribution in panel **b** reveals that the majority of molecules carry a net positive charge of +1 or +2, in keeping with the prevalence of cationic and gemini cationic surfactants in the dataset. Neutral molecules account for a substantial

fraction as well, corresponding to the non-ionic and sugar-based subsets, while negatively charged molecules represent the anionic class.

4.2 Baseline Performance

We first evaluate the baseline generative capabilities of the model without any optimisation objective. This step is essential, as any downstream optimisation is only meaningful if the model is able to generate chemically valid, diverse, and non-trivial molecular structures. To this end, we analyse standard generative metrics such as validity, uniqueness, and novelty, as well as the distribution of generated molecules relative to the training dataset.

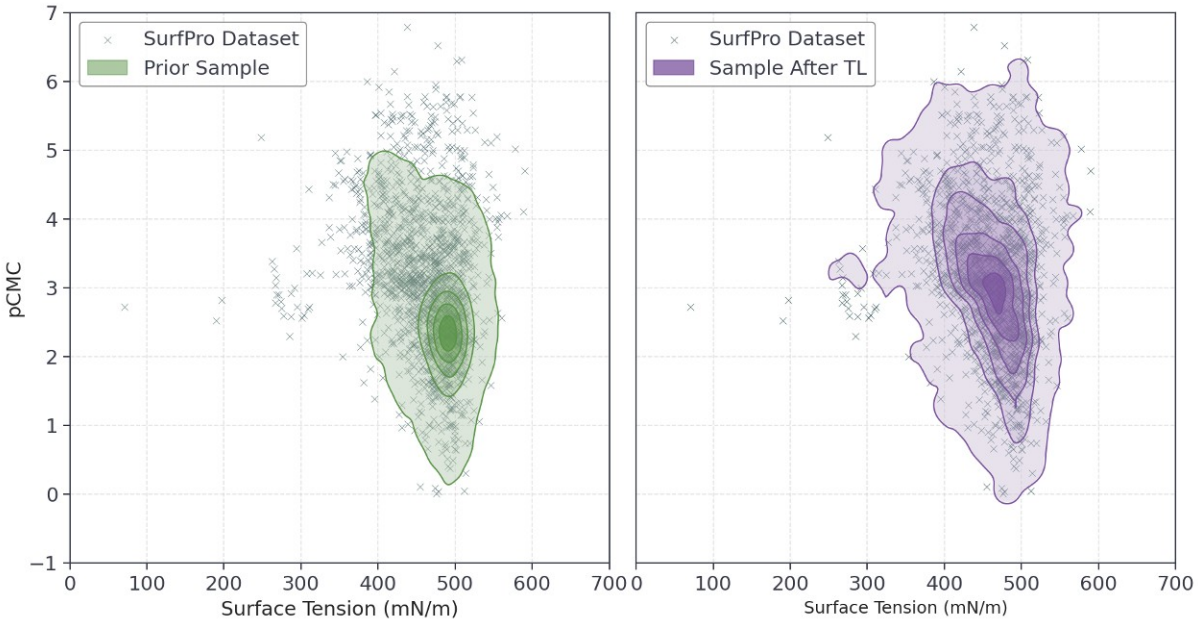

Figure 4.3: Distribution of sampled molcules before and after the prior undergoes transfer learning on the surfpro dataset

We can see in 4.3 that, after the prior has undergone transfer learning, the distribution of sampled molecules from the model clearly shifts to more accurately represent the distribution of the SurfPro dataset. This indicates that the model is now generating molecules with structures more similar to those present in the SurfPro dataset.

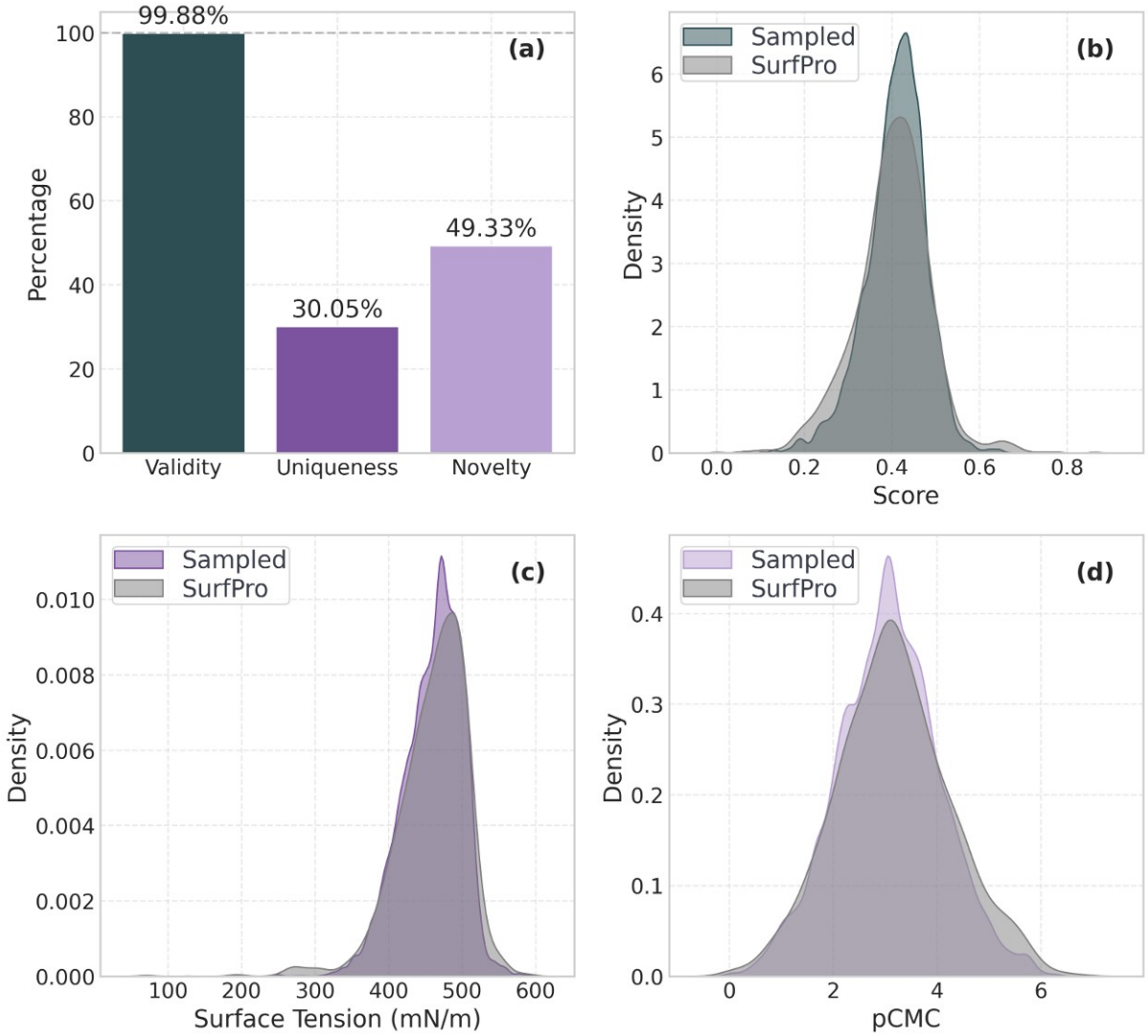

Figure 4.4: Generative quality metrics and property distributions for 30000 unconditionally sampled molecules from the transfer learned model. **(a)** Bar chart of the validity, uniqueness, and novelty of the generated SMILES strings, expressed as percentages. **(b)** Kernel density estimate of the composite score distribution for sampled molecules (teal) overlaid on the SurfPro reference dataset (grey). **(c)** Kernel density estimate of the surface tension distribution (mN/m) for sampled molecules (purple) overlaid on the SurfPro reference dataset (grey). **(d)** Kernel density estimate of the pCMC distribution for sampled molecules (purple) overlaid on the SurfPro reference dataset (grey).

Fig. 4.4 shows the performance of the REINVENT model in unbiased generation. Panel **a** shows the validity, uniqueness, and novelty of the generated SMILES strings. The model generates molecules with nearly 100% validity, 30% uniqueness, and approxiately 50% novelty. Panel **b** shows the distribution of scores of the original SurfPro dataset. Scores have been computed with an equal weighting of pCMC and surface tension as described in Sec. 2.4. The majority of molecules scores between 0.2 and 0.6 with some outliers reaching

as high as 0.7. It should be noted that this score has no immediate physical interpretation due to being a linear combination of normalised pCMC and normalised surface tension, but it provides a useful reference for the evaluation of generated molecules. The model's ability to reproduce pCMC and surface tension distributions is depicted in panels **c** and **d**, respectively. It can be seen that the sampled distributions of both properties closely resemble the distributions found in the SurfPro dataset. In both cases, the sampled distribution is slightly more narrow than the reference distribution. This is expected as all generative models exhibit a tendency to regress to the mean.

4.3 Goal-Directed Optimisation

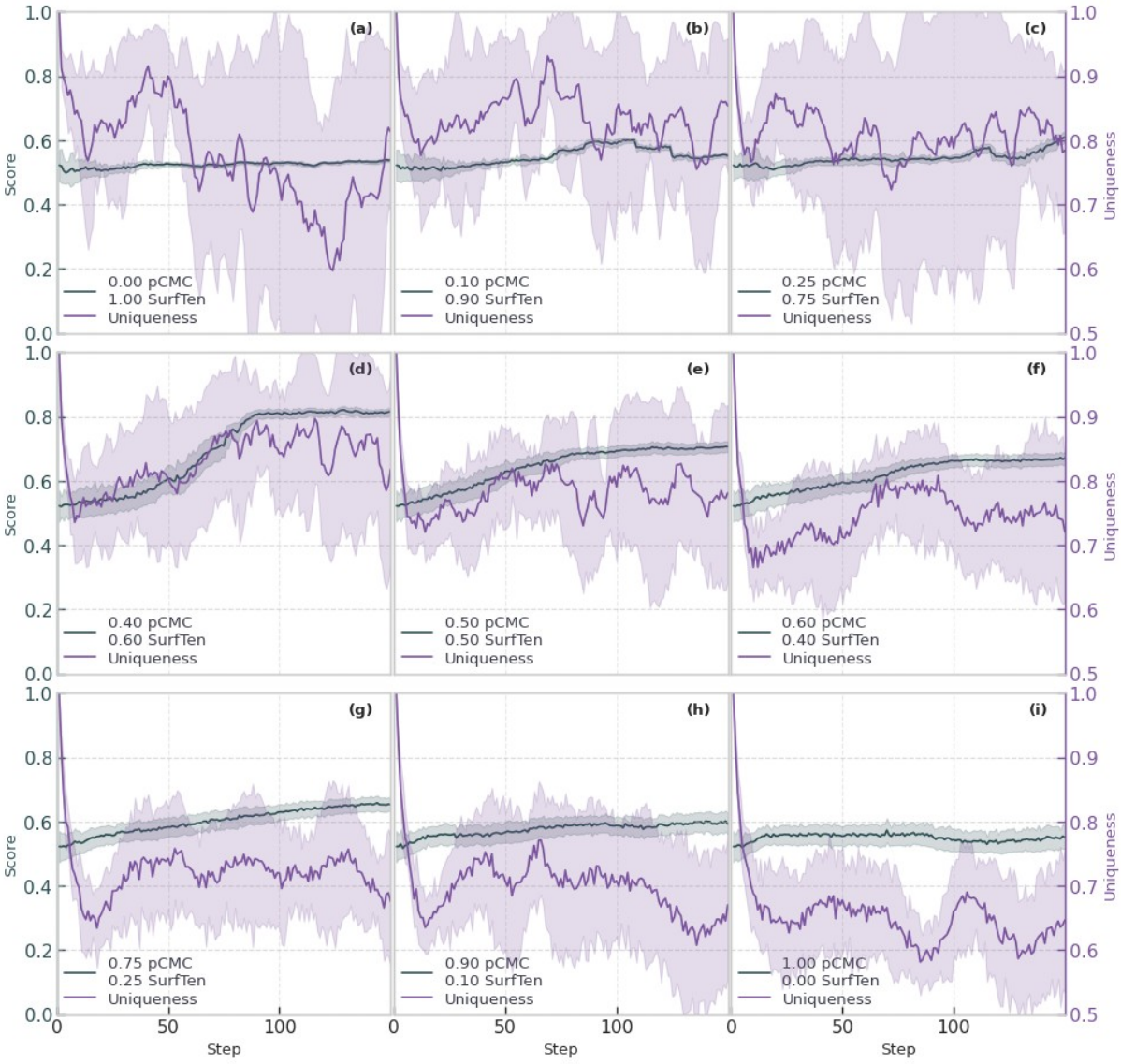

Figure 4.5: Score and uniqueness trajectories over training steps for nine reinforcement learning runs spanning the full pCMCsurface tension weight sweep. Each panel shows the score (left y-axis) and per-step uniqueness (right y-axis) plotted against training step.

Having established that the model can generate valid molecules, we next investigate whether it can be guided towards desirable physicochemical properties using reinforcement learning. To test the model's ability to handle different weightings of pCMC and surface tension, we performed biased generation procedures with different weightings, between a model purely optimising for surface tension and a model purely optimising for pCMC. The score–step curves of all models are shown in Fig. 4.5. Notably, the score does not significantly increase in any of the models in which the weight of the pCMC scoring

function is 25% or less. Similarly, the score does not improve when the contribution of the surface tension scoring function is 10% or less. The best performance is seen for those models which take into account both pCMC and surface tension. Different use cases for surfactants require different weightings between these targets or could involve a different set of surrogate models altogether. In this work, we will continue with an equal weighting between these two components.

However, strong weighting towards a single objective can lead to mode collapse, where diversity decreases and the model focuses on a narrow subset of structures. This highlights the importance of balancing objectives and maintaining exploration. This dynamic is also directly visible in the uniqueness curves overlaid on Fig. 4.5. Models that are strongly weighted towards a single objective show a progressive decline in uniqueness over training steps, while those with a balanced weighting between pCMC and surface tension maintain a higher level of diversity throughout. This confirms that multi-objective scoring acts as a regulariser, preventing the model from collapsing onto a narrow set of high-scoring but structurally redundant molecules.

4.3.1 Advanced Sampling

Multiple methods have been implemented to further push the boundaries of what the model can achieve. A comprehensive comparison of all the different methods is shown in Fig. 4.6. Panel **a** shows the locations of the original SurfPro molecules in the tension-pCMC plane, whereas panel **b** shows the distribution of molecules found during unbiased generation after transfer learning. These distributions confirm the findings detailed in Sec. 4.1 — the unbiased generation mostly reproduces the reference distribution, but it does not recover many of the outliers, especially those with low surface tension. Panels **c** to **e** show the generated distributions for simple runs extremely biased towards pCMC, balanced weighting, and extremely biased towards surface tension, respectively. Notably, a balanced approach explores a much larger region of target space than either of the extremes. The balanced approach generated molecules with significantly lower surface tension than the model taking into account surface tension alone. The same effect can be seen for pCMC, even though it is less pronounced than for surface tension.

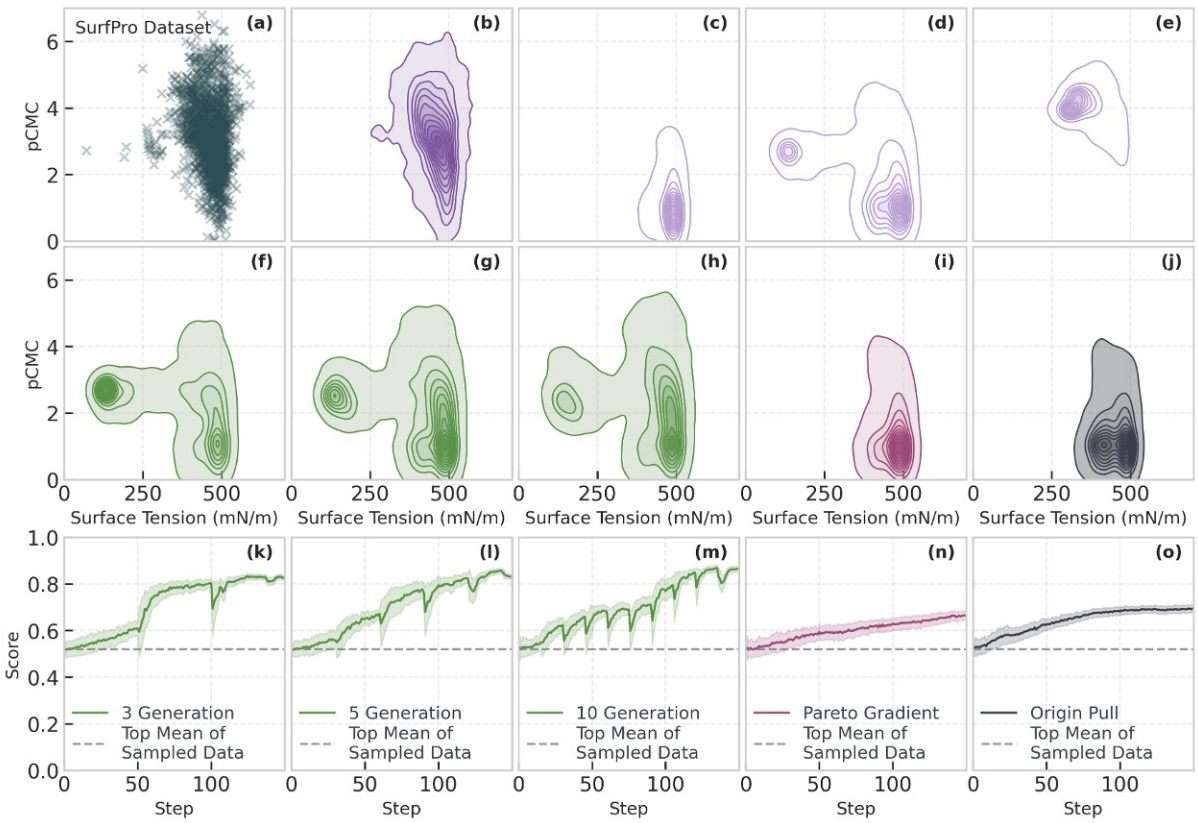

Figure 4.6: Distributions of generated molecules in pCMCsurface tension space and score trajectories over training steps for all optimisation methods. Panels (a)(j) show scatter plots of pCMC versus surface tension (mN/m); panels (k)(o) show score trajectories over training steps, with the solid line indicating the method mean and the grey dashed line indicating the top-decile mean of the unguided sampling baseline. **(a)** SurfPro reference dataset. **(b)** Unguided sampling baseline. **(c–e)** RL-guided generation with different weight balancing. **(f–h)** Iterative retraining with 3, 5 and 10 generations. **(i)** Pareto Gradient method. **(j)** Origin Pull method. **(k–m)** Score trajectory for iterative retraining with 3, 5 and 10 generations. **(n)** Score trajectory for the Pareto Gradient method. **(o)** Score trajectory for the Origin Pull method.

The first of the advanced sampling methods consisted of retraining the generative model on its generated molecules after a fixed number of steps. To keep results comparable, a constant total number of steps was kept while splitting into 3, 5, or 10 generations. Distributions are shown in panels **f** to **h** whereas the evolution of the scoring function is shown in panels **k** to **m**.

Panel **i** shows the distribution of molecules generated using the Pareto gradient method. In contrast to the multi-generation approaches, the Pareto gradient distribution is more compact, concentrating generated molecules in a narrower region of the surface tension-pCMC plane. While this method successfully pushes molecules towards the Pareto front,

it does so with less breadth in its coverage of the objective space compared to the iterative retraining approach. Panel **j** shows the distribution produced by the Origin Pull method, which achieves the lowest hypervolume of all methods compared in this figure. Generated molecules are concentrated in a relatively tight cluster at low pCMC values, suggesting that this method applies a strong directional bias towards one region of target space at the expense of diversity across the full Pareto front.

Panels **n** and **o** show the score trajectories for the Pareto Gradient and Origin Pull methods, respectively. In both cases the score rises steadily over training with the Origin Pull ultimately converging to a plateau. However, the notably lower hypervolume achieved by Origin Pull (1.1328, compared to 1.44 – 1.46 for the best multi-generation runs) highlights the important distinction that a high mean score does not necessarily correspond to broad coverage of the Pareto front. A method can optimise the scoring function effectively while still confining generated molecules to a restricted region of the objective space. This finding underscores the value of using hypervolume as a complementary metric alongside the scoring function when evaluating multi-objective generative methods.

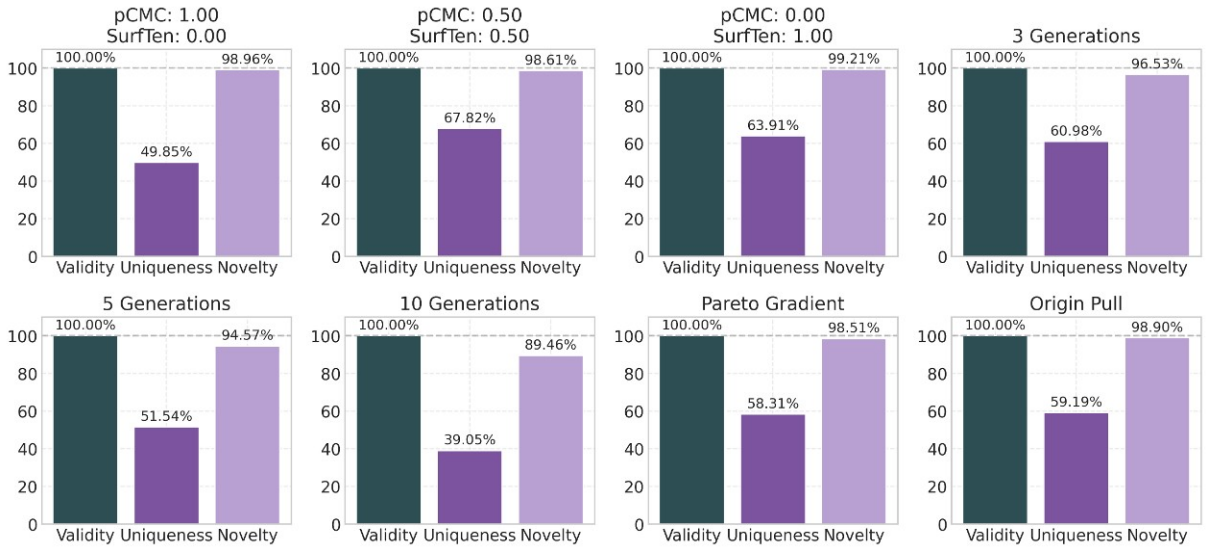

Figure 4.7: Bar charts of validity, uniqueness, and novelty as percentages for each of the eight optimisation methods.

4.3.2 Attribute Analysis

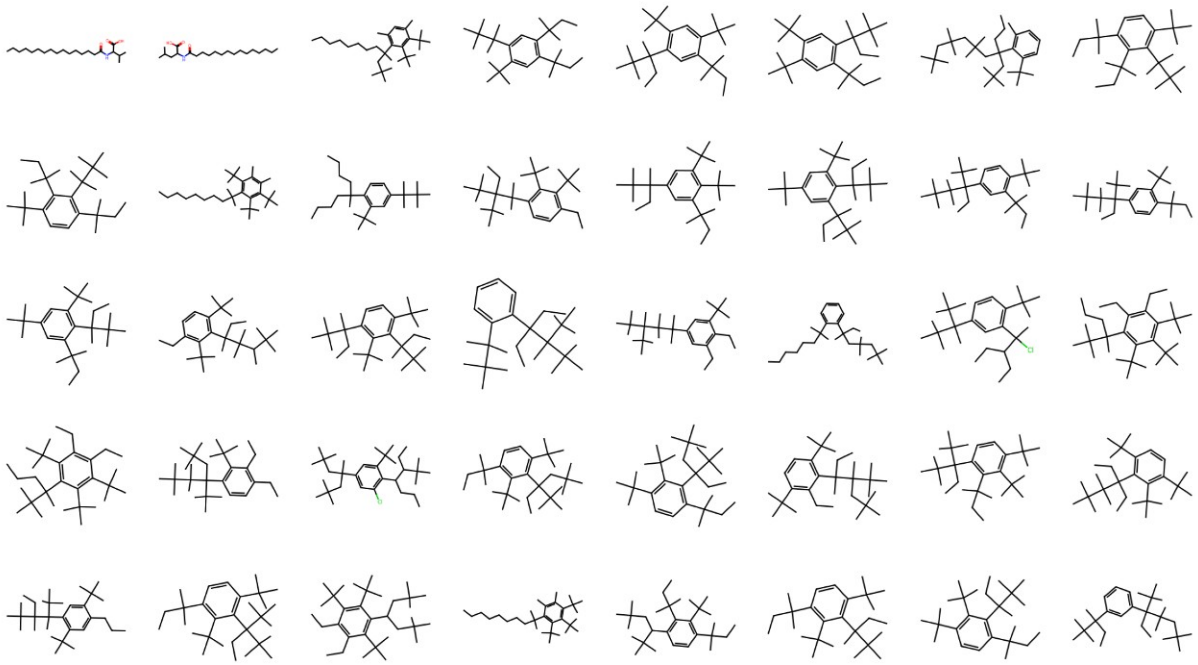

Figure 4.8: Two-dimensional structural drawings of the 40 highest-scoring molecules generated by the iterative retraining method with 5 generations. Serving as a visual example of model convergence towards the hydrocarbon trap.

Visual inspection of the highest-scoring molecules produced by each method reveals a recurring failure mode across several of the optimisation strategies. Rather than generating molecules that resemble genuine surfactants, the model frequently converges on large, heavily branched hydrocarbons that bear no structural resemblance to amphiphilic compounds. These molecules lack any polar head group, contain no heteroatoms, and are therefore incapable of the surface activity that defines a functional surfactant. An illustrative example is shown in Fig. 4.8, where the top 40 molecules generated by the 5-generation retraining method are displayed. Only the two highest-ranked molecules retain a recognisable surfactant structure, a long aliphatic chain attached to a polar, nitrogen- and oxygen-containing head group. The remaining 38 molecules are structurally characterised by polycyclic aromatic cores, typically indene- or fluorene-like bicyclic systems, decorated with multiple tert-butyl substituents, with no polar functionality whatsoever.

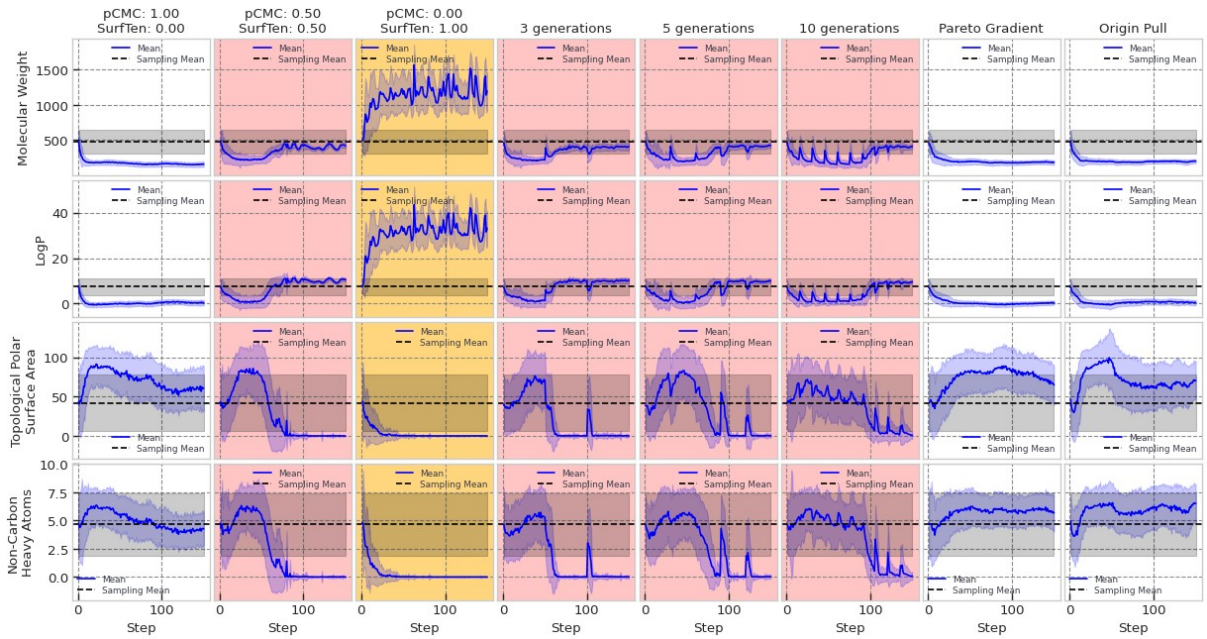

Figure 4.9: Time-series plots of four physicochemical descriptors tracked over the course of RL training across all eight optimisation methods. Rows correspond to molecular weight (top row), MolLogP (second row), topological polar surface area, TPSA (third row), and number of non-carbon heavy atoms (bottom row). Each panel shows the mean descriptor value of generated molecules (solid blue line) relative to the mean of the unbiased sampling baseline (dashed black line), with shaded regions indicating one standard deviation. Columns corresponding to methods exhibiting highly branched like those shown in 4.8 are highlighted in red. Methods exhibiting long, unbranched hydrocarbons are highlighted yellow. The remaining methods are shown on a white background.

To assess the extent and consistency of this behaviour across all methods, a set of physicochemical descriptors was tracked over the course of training and compared against the mean of the unbiased sampling distribution. The results for the four most informative descriptors are shown in Fig. 4.9. The problematic methods, surface tension-only optimisation long with all three iterative retraining variants, are highlighted in red. The balanced weighting, pCMC-only, Pareto gradient, and Origin Pull methods are shown in white and do not exhibit this behaviour to a meaningful extent. Highlighted with black in the plots is the mean values of the top 10% highest scoring molecules from the unbiased sampling as a comparative baseline.

The four descriptors together characterise the trap in full. MolLogP rises to the high end of values sampled from the model, indicative of increased hydrophobicity. TPSA simultaneously collapses to near zero, confirming the complete loss of the polar head group that is essential to surfactant function. The number of non-carbon heavy atoms

in each molecule drops to approximately zero in the problematic methods, demonstrating that the generated structures become pure hydrocarbons with no heteroatoms. Finally, the molecular weight also climbs to the higher end observed from the best molecules in the unbiased sampling. In contrast, these four descriptors remain stable and close to the SurfPro reference distribution throughout training for the non-highlighted methods, confirming that those strategies do not suffer from the same failure mode.

Among the generative retraining variants, the problem is present in all three but appears to be exacerbated by retraining: by iteratively fine-tuning the model on its own problematic outputs, each subsequent generation reinforces and amplifies the deviation from chemically sensible structures.

The surface tension-only weighting (highlighted yellow) presents a closely related but structurally distinct failure mode. As shown in the third column of Fig. 4.9, this method exhibits the most extreme drift of any run in terms of molecular weight and LogP with values substantially exceeding even those of the red-highlighted methods by the end of training. Correspondingly, TPSA and the number of non-carbon atoms converge to near zero, indicating the same complete loss of polar character observed in the retraining methods. However, visual inspection of the top molecules produced by this run reveals that the underlying structures are qualitatively different: rather than compact polycyclic hydrocarbons with tert-butyl clusters, the model produces very long, mostly linear or lightly branched aliphatic chains structures resembling extreme, high-molecular-weight alkanes. Where the generation retraining methods exploit the surface tension surrogate through dense, globular scaffolds, the surface tension-only run exploits it through unbounded chain elongation, apparently having learned that longer hydrophobic chains correlate with lower predicted surface tension. Both modes are chemically invalid as surfactants, the absence of any polar head group makes them incapable of genuine surface activity, but they represent two mechanistically distinct routes to the same failure. This observation suggests that the surface tension surrogate is susceptible to exploitation along multiple dimensions of chemical space, and that neither structural compactness nor chain linearity alone is sufficient to guarantee physically meaningful outputs.

4.4 Post Processing

The filter bounds derived from the SurfPro dataset are reported in Table 4.1. The Branched Carbon Ratio was bounded between 0.0 and 0.16667, reflecting the range of branching character observed across the full breadth of the reference surfactant collection. The MolLogP bounds span a considerably wider interval, from -6.3471 to 21.657, consistent with the diverse lipophilicity profile of a dataset that includes charged gemini cationic, anionic, and sugar-based non-ionic surfactants. The distribution shifts induced by the filter across all methods are visualised in Figure 4.11, which shows the LogP and Branched Carbon Ratio distributions before and after filtering for each method alongside the resulting distribution in the pCMC - surface tension plane.

Branched Carbon Ratio		LogP	
Lower	Upper	Lower	Upper
0.0	0.16667	-6.3471	21.657

Table 4.1: Lower and upper bounds used for post-processing filtering, derived from the range of values observed across the SurfPro dataset, for the Branched Carbon Ratio and MolLogP descriptors.

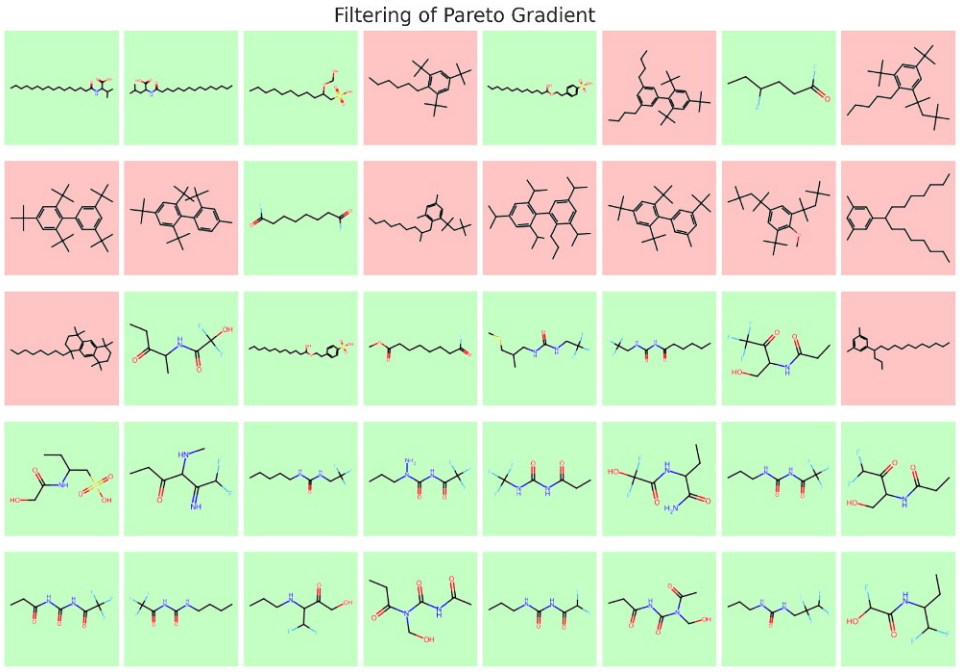

Figure 4.10: Molecules passing the filter are highlighted green while molecules caught by the filter and discarded are highlighted red. This figure shows the top molecules of the *Pareto Gradient* method as a visual example. Further examples can be found in Appendix B.

A visual example is provided in Figure 4.10 where we can see that the high-scoring hydrocarbons, bearing no resemblance to surfactants, are being successfully identified and eliminated.

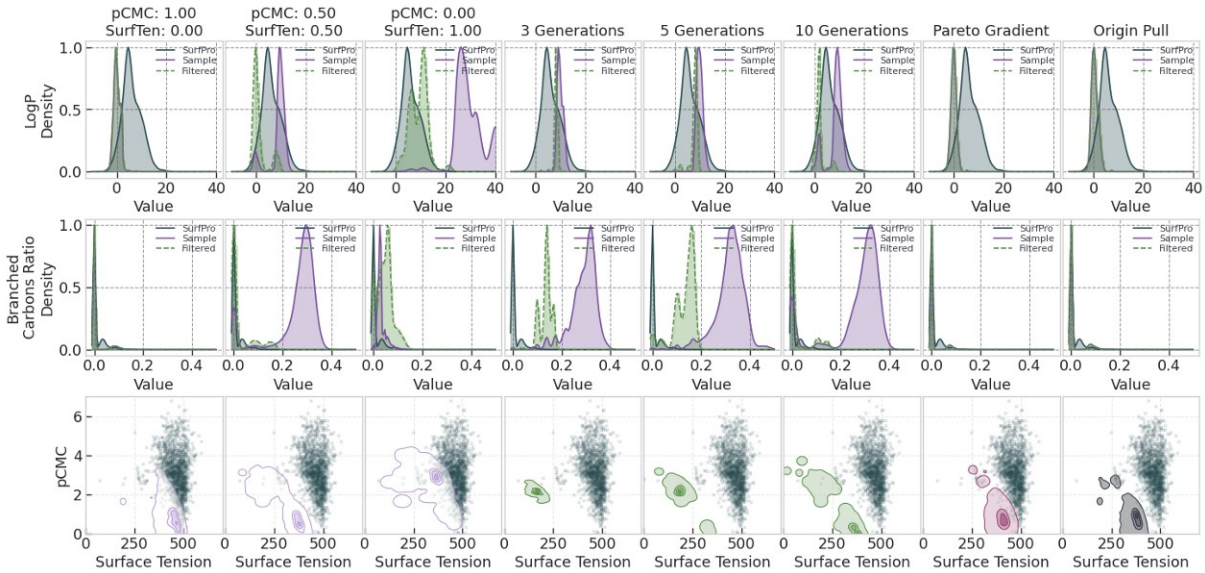

Figure 4.11: Distribution shifts induced by post-processing filtering for all eight optimisation methods, with each method occupying one column. The top row shows kernel density estimates of the LogP distribution for the SurfPro reference (grey), the pre-filter generated sample (purple), and the post-filter generated sample (dashed green). The middle row shows the corresponding Branched Carbon Ratio distributions in the same format. The bottom row shows scatter plots of pCMC versus surface tension (mN/m) for the post-filter generated molecules (coloured) overlaid on the SurfPro reference (grey).

Method	# Molecules in top 10%	# Caught by filter	# Remaining	Remaining (%)
pCMC: 1.00 SurfTen: 0.00	36853	81	36772	99.78%
pCMC: 0.50 SurfTen: 0.50	37335	32393	4942	13.24%
pCMC: 0.00 SurfTen: 1.00	37023	36323	700	1.89%
3 Generations	36710	35278	1432	3.90%
5 Generations	36476	35375	1101	3.02%
10 Generations	35612	29967	5645	15.85%
Pareto Gradient	37207	66	37141	99.82%
Origin Pull	36894	169	36725	99.54%

Table 4.2: Post-processing filter statistics for each of the eight optimisation methods, reporting the total number of molecules in the top 10% by score, the number removed by the physicochemical filter, the number of molecules retained, and the retention rate as a percentage of the pre-filter total.

The retention statistics reported in Table 4.2 reveal a striking divergence between methods, one that closely mirrors the failure modes identified in the attribute analysis of Section 4.3.2. The three methods that exhibited the most severe structural degradation during training are also the most aggressively filtered here. The surface tension-only weighting (pCMC: 0.00–SurfTen: 1.00) retains only 700 of 37 023 top-decile molecules, a survival rate of 1.89%, confirming that the extreme chain elongation and near-total loss of polar character documented for this method produces structures that fall almost entirely outside the LogP range of the SurfPro reference. The iterative retraining variants show a similarly severe attrition: *3 Generations* and *5 Generations* retain 3.90% and 3.02% of molecules respectively, while *10 Generations* fares marginally better at 15.85%, though even this residual fraction represents only a small minority of the total generated set. The iterative self-reinforcement of chemically degenerate outputs in these methods, discussed in Section 4.3.2, is directly reflected in the near-total collapse of their candidate pools under the filter.

In contrast, the pCMC-only weighting (pCMC: 1.00–SurfTen: 0.00), *Pareto Gradient*, and *Origin Pull* methods pass the filter with almost no attrition, retaining 99.78%, 99.82%, and 99.54% of their top-decile molecules respectively. This is consistent with the finding that these methods do not exhibit meaningful descriptor drift over the course of training, and the generated molecules therefore remain broadly within the physicochemical envelope of known surfactants. The balanced weighting (pCMC: 0.50–SurfTen: 0.50) occupies an intermediate position, retaining 13.24% of molecules. Although this method was not among those flagged as severely problematic in the attribute analysis, the partial influence of the surface tension scoring component is evidently sufficient to push a large fraction of generated molecules outside the SurfPro LogP bounds, even if the structural collapse is less extreme than in the surface tension-only case. Taken together, the post-processing filter provides a concise and quantitative confirmation of which optimisation strategies produce chemically viable candidates: those that avoid strong, unidirectional pressure towards surface tension minimisation preserve the physicochemical character of real surfactants, while those that apply it, whether directly or through iterative self-training, generate large volumes of molecules that cannot pass even a minimal viability screen.

5

Conclusion

5.1 Discussion

The results presented in this thesis demonstrate that a reinforcement learning-guided generative model can be effectively applied to the inverse design of surfactant molecules. In this section, we interpret the key findings, reflect on the behaviour of the framework, discuss its limitations, and consider the broader implications of the work.

5.1.1 Improvement of the Pareto Front

The central question motivating this work was whether the proposed framework could generate novel surfactant molecules that improve upon the Pareto front of critical micelle concentration (pCMC) and surface tension defined by the SurfPro dataset. The results indicate that this is indeed achievable. Following reinforcement learning, a meaningful fraction of generated molecules extended beyond the Pareto front of the training data, suggesting that the model is capable of exploring regions of chemical space not represented in the original dataset. While the precise degree of improvement is difficult to quantify, given the reliance on surrogate model predictions rather than experimental measurements, the directional movement of the generated distribution towards and beyond the known Pareto front is a consistent and reproducible finding across different scoring function configurations. This represents a meaningful result, as it demonstrates that the generative model is not merely interpolating within the training distribution but is actively proposing candidate structures with potentially superior properties.

5.1.2 Limitations of the Scoring Functions

A recurring observation upon visual inspection of the top-scoring generated molecules is that many do not exhibit the clear amphiphilic headtail architecture that is the defining structural hallmark of a functional surfactant. A genuine surfactant molecule requires a hydrophilic head group and a hydrophobic tail. Without this separation a molecule will not adsorb at an airwater or oilwater interface, it will not lower surface tension through the conventional mechanism, and it will not self-assemble into micelles at a meaningful concentration. Despite this, many such structures score favourably under the surrogate-model-based scoring functions employed in this work.

This is best understood as a limitation of the scoring functions rather than a failure of the generative model. The surrogate models predict pCMC and surface tension from a set of molecular descriptors, and it is possible for a molecule to occupy a region of descriptor space that happens to be associated with favourable predicted values without that molecule possessing the structural prerequisites for real surfactant behaviour. The generative model itself is functioning correctly, it is producing chemically valid, novel structures that optimise the objective it has been given. The issue is that the objective does not fully capture what it means for a molecule to be a surfactant.

Addressing this gap would require augmenting the reward signal with structural constraints. One approach would be to penalise molecules that lack a clear hydrophilichydrophobic separation, for example by incorporating a descriptor that quantifies amphiphilicity directly. Another would be to restrict the vocabulary of generated structures to those containing known surfactant head groups, either through hard constraints on the generative model or through an additional term in the scoring function. Incorporating such constraints in future work would improve the practical relevance of the generated candidates and reduce the rate at which high-scoring molecules are dismissed upon visual inspection.

5.1.3 Failure modes

The reinforcement learning experiments revealed two distinct but related failure modes, both of which arise as consequences of excessively narrow or unbalanced optimisation pressure, and both of which carry important implications for the design of future generative

frameworks in this domain.

The first is mode collapse under strongly biased objective weighting. When the scoring function placed strong weight on a single objective, the diversity of generated molecules decreased substantially, with the model concentrating on a narrow subset of structures that scored highly on the dominant objective while neglecting the broader chemical space. This behaviour is a well-known consequence of the explorationexploitation trade-off inherent to reinforcement learning: when the reward signal is dominated by a single objective, the policy gradient updates to consistently reinforce a small set of high-scoring structural motifs, causing the model to assign increasing probability mass to those motifs at the expense of exploration. The implication for scoring function design is clear, maintaining a balance between objectives is not merely a matter of preference but a practical necessity for preserving chemical diversity. This is directly corroborated by the uniqueness-over-steps panel in Fig. 4.5, which shows that configurations dominated by a single objective exhibit a more pronounced and sustained drop in per-step uniqueness relative to balanced configurations. In future work, explicitly incorporating diversity or entropy bonuses into the reward signal could help mitigate this effect and allow the model to more thoroughly explore the Pareto-optimal region.

The second failure mode, which is referred to here as the *hydrocarbon trap*, is in many respects a more severe manifestation of the same underlying dynamic. Rather than simply losing diversity, certain configurations, most notably the iterative retraining variants and the surface tension-only weighting, converge on structurally degenerate molecules that score highly according to the surrogate models while bearing no resemblance to functional surfactants. This is an instance of a broader phenomenon known as reward hacking [21], which occurs when an agent learns to exploit weaknesses in the proxy reward signal rather than genuinely optimising the underlying objective. In the iterative retraining methods, the model converges on compact polycyclic hydrocarbons decorated with multiple tert-butyl substituents, entirely hydrophobic structures with no polar character whatsoever. In the surface tension-only weighting, the model instead produces very long, predominantly linear, high-molecular-weight aliphatic chains. Both outcomes represent the surrogate models being deceived by molecules that occupy regions of descriptor space associated with favourable predicted values, while being physically incapable of genuine surfactant behaviour: without a hydrophilic head group, a molecule cannot adsorb at an air-water

interface, cannot lower surface tension through the conventional mechanism, and cannot self-assemble into micelles at any meaningful concentration.

The convergence of the surface tension-only run on extended aliphatic chains is perhaps the less surprising of the two, long hydrophobic molecules are chemically associated with low water solubility, a property the surrogate may plausibly correlate with low predicted surface tension, but the polycyclic structures produced by the iterative retraining methods represent a more unexpected exploitation route. The underlying vulnerability in both cases is most plausibly attributed to a shared limitation of the descriptor set. The approximately 50 RDKit descriptors used as surrogate model inputs do not directly encode amphiphilicity as a concept, meaning that a molecule can achieve favourable descriptor values while bearing no structural resemblance to a genuine surfactant. The post-processing filter introduced in Section 3.6 provides a practical means of eliminating the most severely degenerate outputs after the fact; however, a more principled long-term solution would be to incorporate amphiphilicity-aware constraints directly into the scoring function, preventing the trap from being entered in the first place. Notably, the iterative retraining approach appears particularly susceptible to this failure mode because self-reinforcement amplifies the problem across generations: by repeatedly fine-tuning the model on its own degenerate outputs, each generation embeds the problematic structural patterns more deeply into the model's learned distribution, making course-correction progressively more difficult.

5.1.4 Interpreting Uniqueness in Unconditional Generation

The low uniqueness score of approximately 30% observed during unconditional generation warrants some discussion, though it should not be interpreted as a fundamental flaw of the model. In the absence of any scoring function, the generative model tends to produce repeated structures, both duplicates of training molecules and self-repetition across the generated batch. This is expected behaviour for a model that has been fine-tuned on a relatively small and structurally coherent dataset such as SurfPro. The metric is reported primarily to establish a baseline characterisation of the model's unconditional behaviour, and the important question is how uniqueness evolves once reinforcement learning is applied. This is directly addressed by the uniqueness-over-steps panel in Fig. 4.5. The plot shows that uniqueness drops across all weight configurations immediately at the onset of

RL training, which is expected: in the very first steps every generated molecule is new relative to the batch, so uniqueness starts high and then falls as the model begins to converge. More informative is the long-run trajectory: configurations with balanced weightings between pCMC and surface tension maintain meaningfully higher per-step uniqueness throughout training, whereas configurations in which nearly all weight is placed on a single objective, such as pCMC or surface tension alone, show a lower sustained level of uniqueness. This provides direct empirical support for the mode collapse observations discussed above and confirms that maintaining a balance between objectives is an effective practical strategy for preserving chemical diversity during RL-guided generation.

5.1.5 Dataset Size as a Bottleneck

A significant limitation of the overall framework is the small size of the SurfPro dataset, which contains approximately 1550 molecules after preprocessing. This constrains the framework in two important ways. First, it limits the chemical vocabulary available during transfer learning, causing the fine-tuned model to stay relatively close to the structural motifs represented in the training data, even after reinforcement learning is applied. This may explain why the generated molecules, while able to push beyond the Pareto front, tend to cluster in property space rather than spreading uniformly across it. Second, the small dataset size affects the reliability of the surrogate models, which are used as proxies for experimental measurements throughout the optimisation process. Surrogate models trained on limited data are more likely to exhibit extrapolation errors in unexplored regions of chemical space precisely the regions the model is trying to discover. Expanding the dataset through additional literature curation or high-throughput computational screening would likely improve both the diversity of generated structures and the reliability of property predictions.

5.1.6 Implications for PFAS Replacement

More broadly, this work contributes to the emerging field of AI-assisted molecular design for sustainable chemistry. The ability to computationally propose novel surfactant candidates with improved pCMC and surface tension properties, in a sample-efficient manner, has clear potential relevance to the challenge of replacing per- and polyfluoroalkyl substances (PFAS). The framework developed here demonstrates that the computational

component of this challenge is tractable the generative model can propose genuinely novel candidates that outperform the known dataset on the target properties. However, realising this potential in practice will require close collaboration between computational and experimental researchers to validate, synthesise, and test the most promising candidates identified by the model.

5.2 Summary and future work

This thesis set out to investigate whether a reinforcement learning-guided generative framework could be applied to the inverse design of surfactant molecules with improved physicochemical properties. Starting from a pretrained generative language model adapted to the surfactant domain through transfer learning on the SurfPro dataset, three reinforcement learning strategies were developed, applied, and evaluated against both an unguided sampling baseline and the known Pareto front defined by the reference data. The results provide optimistic answers to the research questions posed in Chapter 1.

With respect to the first, how a generative model can be guided towards industrially relevant surfactant candidates, the experiments demonstrate that reinforcement learning with a carefully designed multi-objective scoring function is a functional approach, capable of reproducibly shifting the generated distribution towards improved regions of property space.

The second question, concerning the most suitable reward function for this optimisation problem, does not admit a single definitive answer, however, the results consistently show that balanced multi-objective scoring outperforms single-objective weighting both in terms of chemical diversity and in avoiding structural collapse, with Pareto-based reward signals offering particular advantages for Pareto front coverage.

As for the third research question, whether the framework can generate surfactants that outperform the current state of the art, the answer is a conditional yes. Molecules that extend beyond the Pareto front of the SurfPro dataset were consistently produced across multiple optimisation configurations, though this conclusion rests on the reliability of the surrogate models used for evaluation and remains to be confirmed experimentally.

What this thesis ultimately demonstrates is a proof of concept. The framework is capable

of navigating surfactant chemical space in a principled and goal-directed manner, and the candidates it produces represent theoretically plausible leads for future experimental investigation. At the same time, the failure modes uncovered, most notably the hydrocarbon trap, make clear that the current framework requires further refinement before it can reliably and systematically produce candidates that are not only high-scoring but also structurally sound.

Improvements to the surrogate models, the incorporation of amphiphilicity-aware constraints into the scoring function, and expansion of the training dataset all represent concrete and well-motivated directions for future work. As one formulation of the key takeaway: the work has demonstrated that there is genuine potential in this approach, but also that there are real limits to what the current framework can achieve on its own. It remains for future work to build on these findings, refine the methodology, and ultimately apply it to more concrete problems with more strictly defined requirements, with the ultimate goal of producing a shortlist of novel, synthesisable surfactant candidates worthy of experimental testing and, perhaps, a new standard of performance.

Bibliography

[1] United States EPA, *PFAS Explained*, https://www.epa.gov/pfas/pfas-explained, Accessed: 2025-12-11, 2025.

[2] D. J. McClements, "Encapsulation, protection, and release of hydrophilic active components: Potential and limitations of colloidal delivery systems," *Advances in Colloid and Interface Science*, vol. 240, pp. 31–59, 2017, ISSN: 0001-8686. DOI: 10.1016/j.cis.2016.12.002.

[3] M. J. Rosen and J. T. Kunjappu, *Surfactants and Interfacial Phenomena*, 4th ed. Hoboken, NJ: John Wiley & Sons, 2012, ISBN: 978-0-470-54194-4.

[4] S. O. Badmus, H. K. Amusa, T. A. Oyehan, and T. A. Saleh, "Environmental risks and toxicity of surfactants: Overview of analysis, assessment, and remediation techniques," *Environ Sci Pollut Res Int*, vol. 28, no. 44, pp. 62 085–62 104, 2021, ISSN: 0944-1344. DOI: 10.1007/s11356-021-16483-w. Accessed: May 27, 2026. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8480275/.

[5] R. F. Nunes and A. C. S. C. Teixeira, "An overview on surfactants as pollutants of concern: Occurrence, impacts and persulfate-based remediation technologies," *Chemosphere*, vol. 300, p. 134 507, Aug. 2022, ISSN: 0045-6535. DOI: 10.1016/j.chemosphere.2022.134507. Accessed: May 27, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045653522010001.

[6] J.-L. Reymond, "The chemical space project," *Accounts of Chemical Research*, vol. 48, no. 3, pp. 722–730, 2015, PMID: 25687211. DOI: 10.1021/ar500432k. eprint: https://doi.org/10.1021/ar500432k. [Online]. Available: https://doi.org/10.1021/ar500432k.

[7] R. Gómez-Bombarelli et al., "Automatic chemical design using a data-driven continuous representation of molecules," *CoRR*, vol. abs/1610.02415, 2016. arXiv: 1610.02415. [Online]. Available: http://arxiv.org/abs/1610.02415.

[8] Y. Du et al., "Machine learning-aided generative molecular design," en, *Nat Mach Intell*, vol. 6, no. 6, pp. 589–604, Jun. 2024, ISSN: 2522-5839. DOI: 10.1038/s42256-024-00843-5. Accessed: May 27, 2026. [Online]. Available: https://www.nature.com/articles/s42256-024-00843-5.

[9] Y. Li, J. Pei, and L. Lai, "Structure-based de novo drug design using 3d deep generative models," *Chemical Science*, vol. 12, no. 41, pp. 13 664–13 675, Oct. 27, 2021, ISSN: 2041-6539. DOI: 10.1039/D1SC04444C. Accessed: May 27, 2026. [Online]. Available: https://pubs.rsc.org/en/content/articlelanding/2021/sc/d1sc04444c.

[10] X. Zeng et al., "Deep generative molecular design reshapes drug discovery," *Cell Reports Medicine*, vol. 3, no. 12, p. 100 794, Oct. 27, 2022, ISSN: 2666-3791. DOI: 10.1016/j.xcrm.2022.100794. Accessed: May 27, 2026. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9797947/.

[11] R. L. van den Broek, S. Patel, G. J. P. van Westen, W. Jespers, and W. Sherman, "In search of beautiful molecules: A perspective on generative modeling for drug design," *Journal of Chemical Information and Modeling*, vol. 65, no. 18, pp. 9383–9397, Sep. 22, 2025, ISSN: 1549-9596. DOI: 10.1021/acs.jcim.5c01203. Accessed: May 27, 2026. [Online]. Available: https://doi.org/10.1021/acs.jcim.5c01203.

[12] M. Nnadili, A. N. Okafor, T. Olayiwola, D. Akinpelu, R. Kumar, and J. A. Romagnoli, "Surfactant-specific AI-driven molecular design: Integrating generative models, predictive modeling, and reinforcement learning for tailored surfactant synthesis," *Industrial & Engineering Chemistry Research*, vol. 63, no. 14, pp. 6313–6324, Apr. 10, 2024, ISSN: 0888-5885. DOI: 10.1021/acs.iecr.4c00401. Accessed: Feb. 26, 2026. [Online]. Available: https://doi.org/10.1021/acs.iecr.4c00401.

[13] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, 1988. DOI: 10.1021/ci00057a005. eprint: https://doi.org/10.1021/ci00057a005. [Online]. Available: https://doi.org/10.1021/ci00057a005.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, Nov. 1997. DOI: 10.1162/neco.1997.9.8.1735.

[16] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd. MIT Press, 2018. [Online]. Available: http://incompleteideas.net/book/the-book-2nd.html.

[17] B. Settles, *Active Learning*. Jun. 2012, vol. 6. DOI: 10.2200/S00429ED1V01Y201207AIM018.

[18] H. H. Loeffler et al., "Reinvent 4: Modern ai-driven generative molecule design," *Journal of Cheminformatics*, vol. Vol.16, no. 20, pp. 1–16, 2024, ISSN: 1758-2946. DOI: 10.1186/s13321-024-00812-5. [Online]. Available: https://doi.org/10.1186/s13321-024-00812-5.

[19] R. Beckmann et al., "A unified experimental-simulation dataset and surrogate models for surfactant property prediction," *ChemRxiv*, vol. 2026, no. 0506, 2026. DOI: 10.26434/chemrxiv.15002790/v1. eprint: https://chemrxiv.org/doi/pdf/10.26434/chemrxiv.15002790/v1. [Online]. Available: https://chemrxiv.org/doi/abs/10.26434/chemrxiv.15002790/v1.

[20] G. Landrum et al., *Rdkit/rdkit: 2026_03_2 (q1 2026) release*, version Release_2026_03_2, Apr. 2026. DOI: 10.5281/zenodo.19922430. [Online]. Available: https://doi.org/10.5281/zenodo.19922430.

[21] T. Yoshizawa, S. Ishida, T. Sato, M. Ohta, T. Honma, and K. Terayama, "A data-driven generative strategy to avoid reward hacking in multi-objective molecular design," *Nature Communications*, vol. 16, no. 1, p. 2409, Mar. 2025, ISSN: 2041-1723. DOI: 10.1038/s41467-025-57582-3. [Online]. Available: https://doi.org/10.1038/s41467-025-57582-3.

[22] S. L. Hödl, L. Hermans, P. F. J. Dankloff, A. Piruska, W. T. S. Huck, and W. E. Robinson, "Surfpro a curated database and predictive model of experimental properties of surfactants," *Digital Discovery*, vol. Vol.4, pp. 1176–1187, 5 2025. DOI: 10.1039/D4DD00393D. [Online]. Available: http://dx.doi.org/10.1039/D4DD00393D.

A

Appendix 1

A.1 Code and Data Availability

The code, models and datasets used can be found at:

`https://github.com/hannesohman/REINVENT4Surfactants`

I would like to express my gratitude to Hannes Loeffler, the lead developer on REINVENT4 for providing the prior trained on PubChem.

`https://github.com/MolecularAI/REINVENT4/discussions/316`

`https://drive.google.com/file/d/1rOwWruKtWHo5jyJ7s1-4guOSg_IiX_OK/view?usp=sharing`

B

Appendix 2

B.1 Post-hoc filtering examples

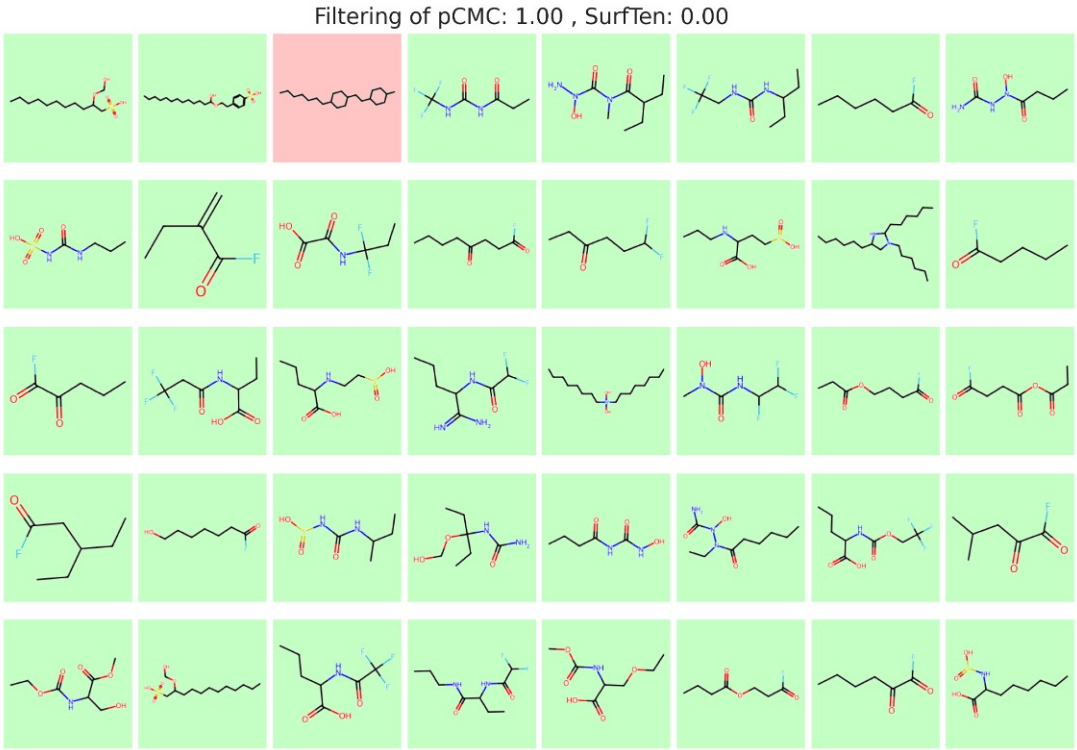

Figure B.1: Filter effect on pCMC: 1.00 , Surface Tension: 0.00

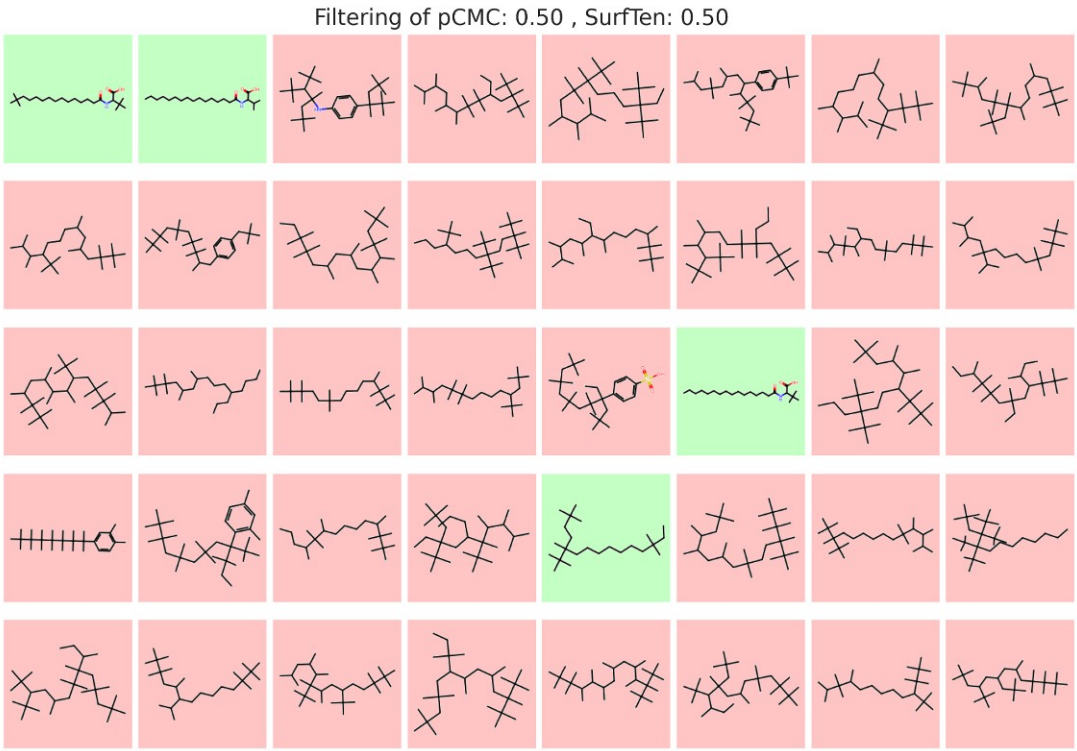

Figure B.2: Filter effect on pCMC: 0.50 , Surface Tension: 0.50

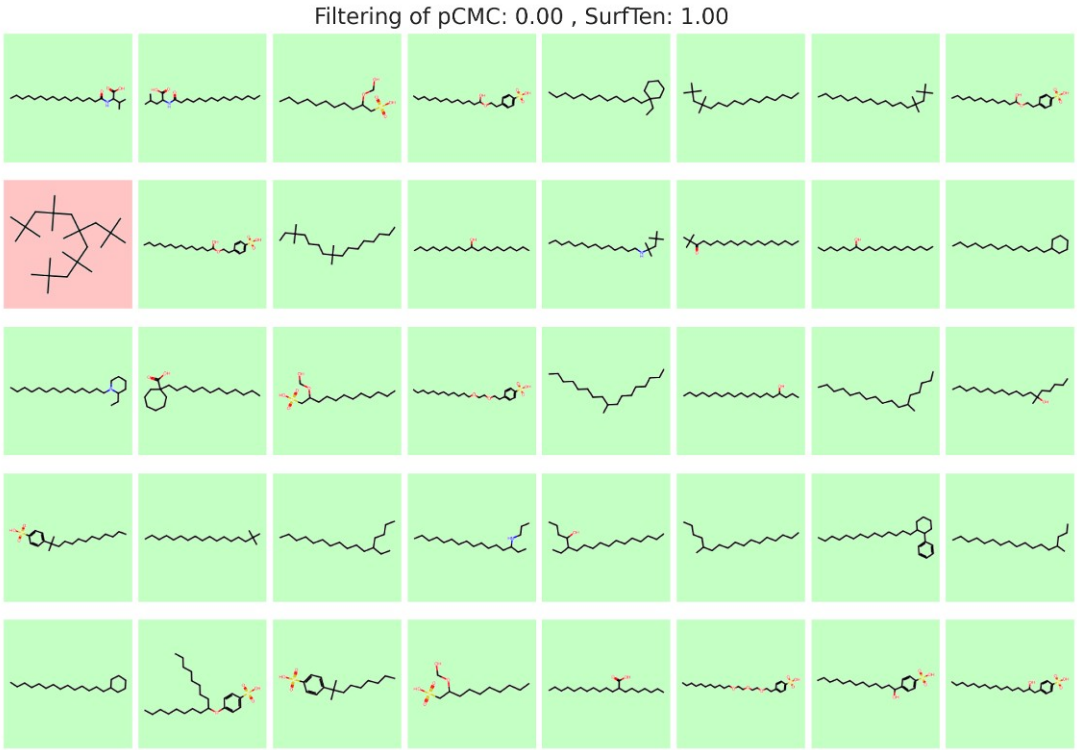

Figure B.3: Filter effect on pCMC: 0.00 , Surface Tension: 1.00

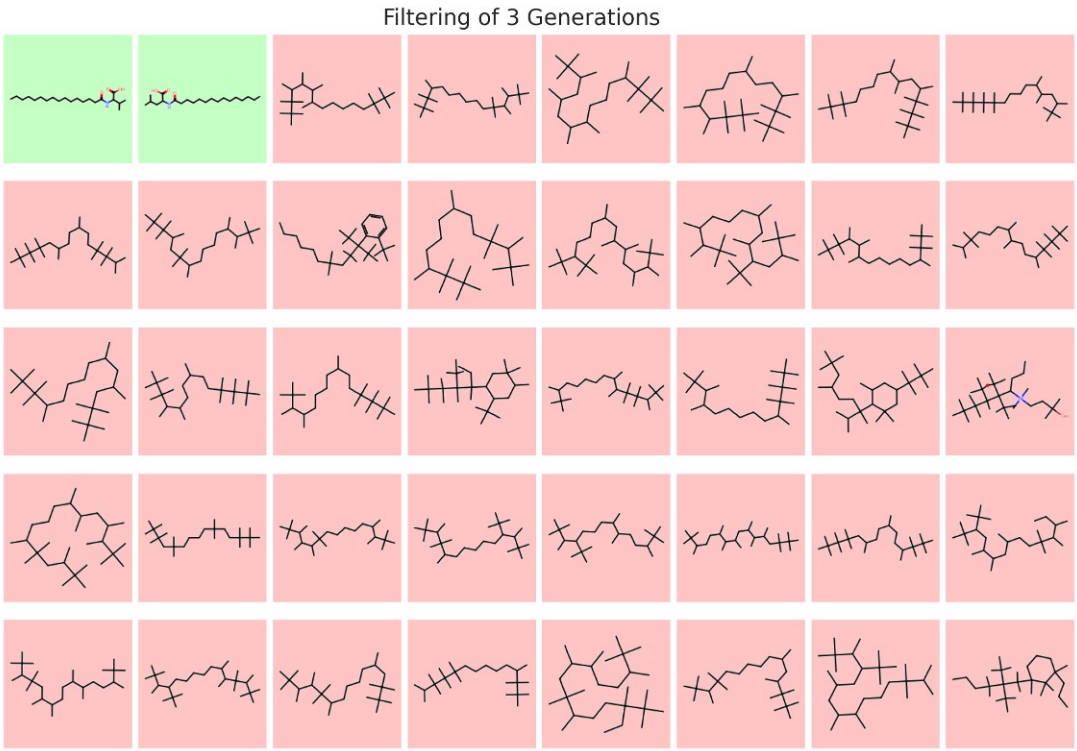

Figure B.4: Filter effect on Iterative Retraining (3 Generations)

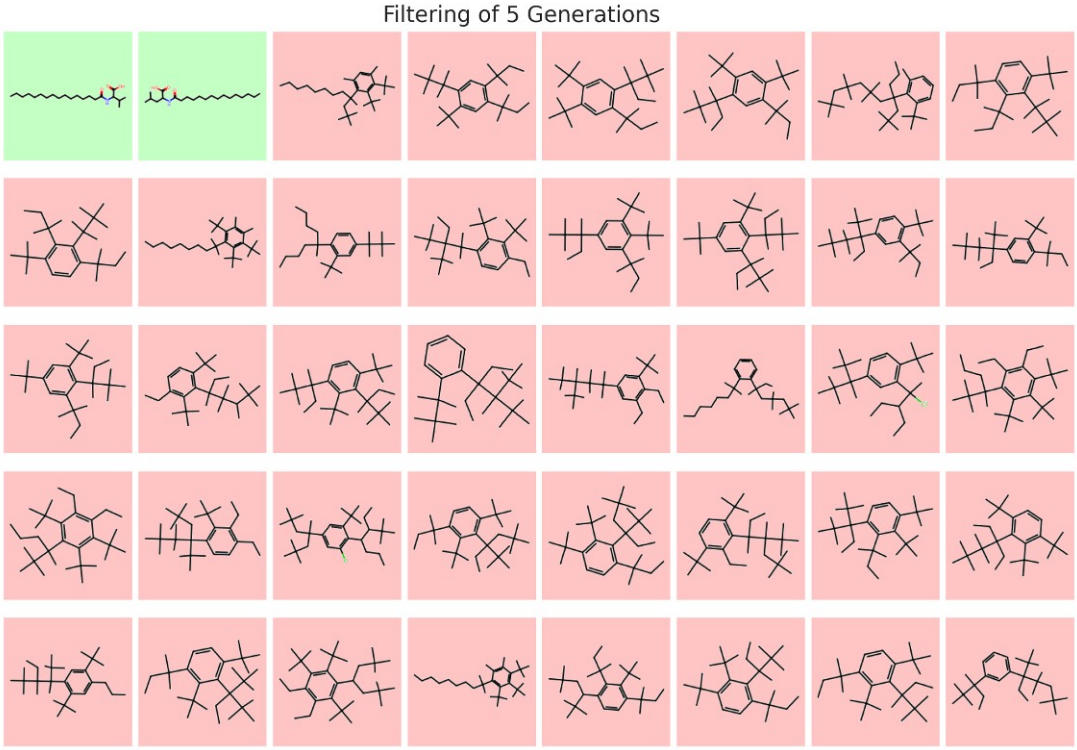

Figure B.5: Filter effect on Iterative Retraining (5 Generations)

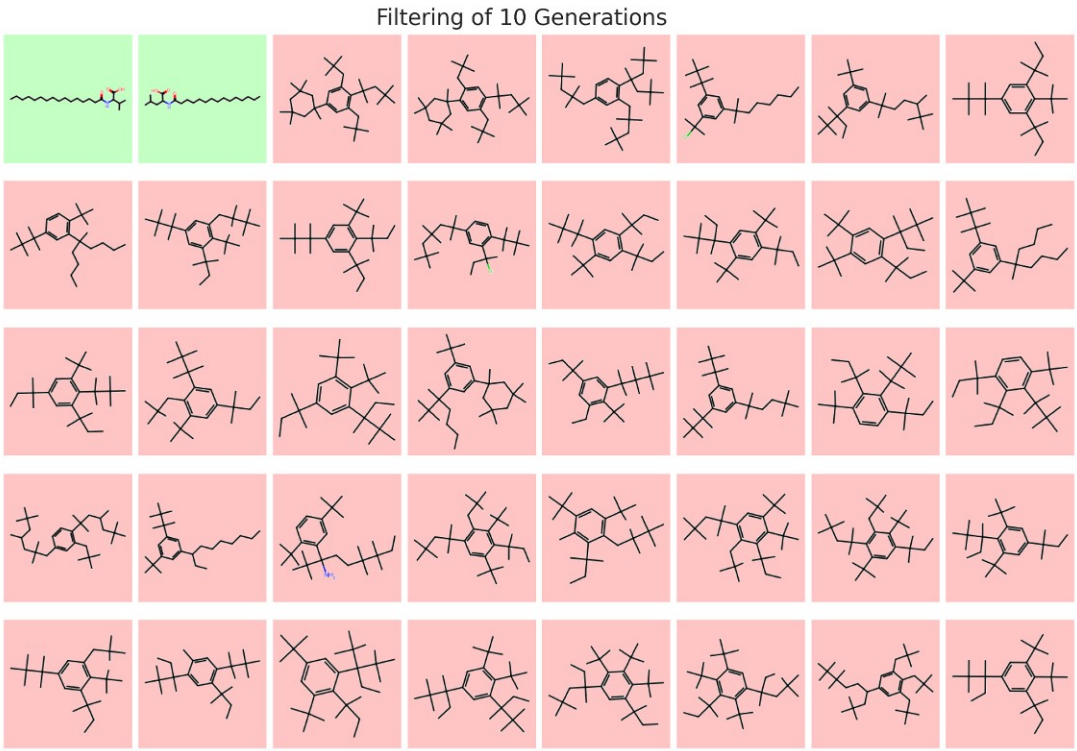

Figure B.6: Filter effect on Iterative Retraining (10 Generations)

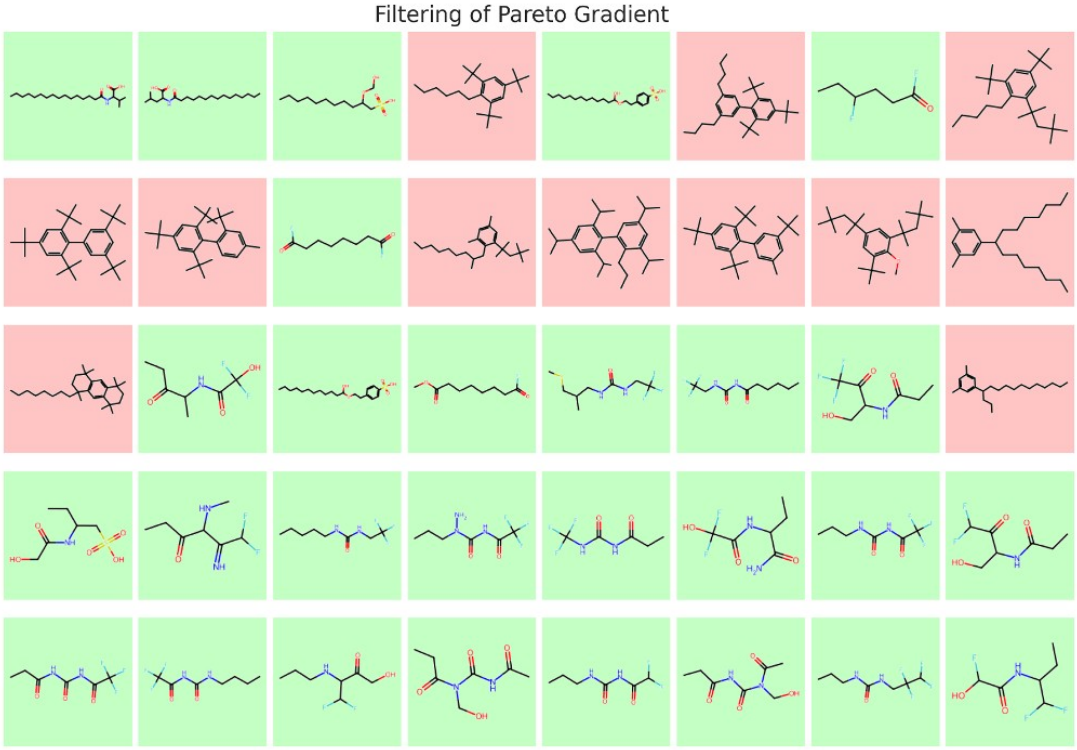

Figure B.7: Filter effect on Pareto Gradient

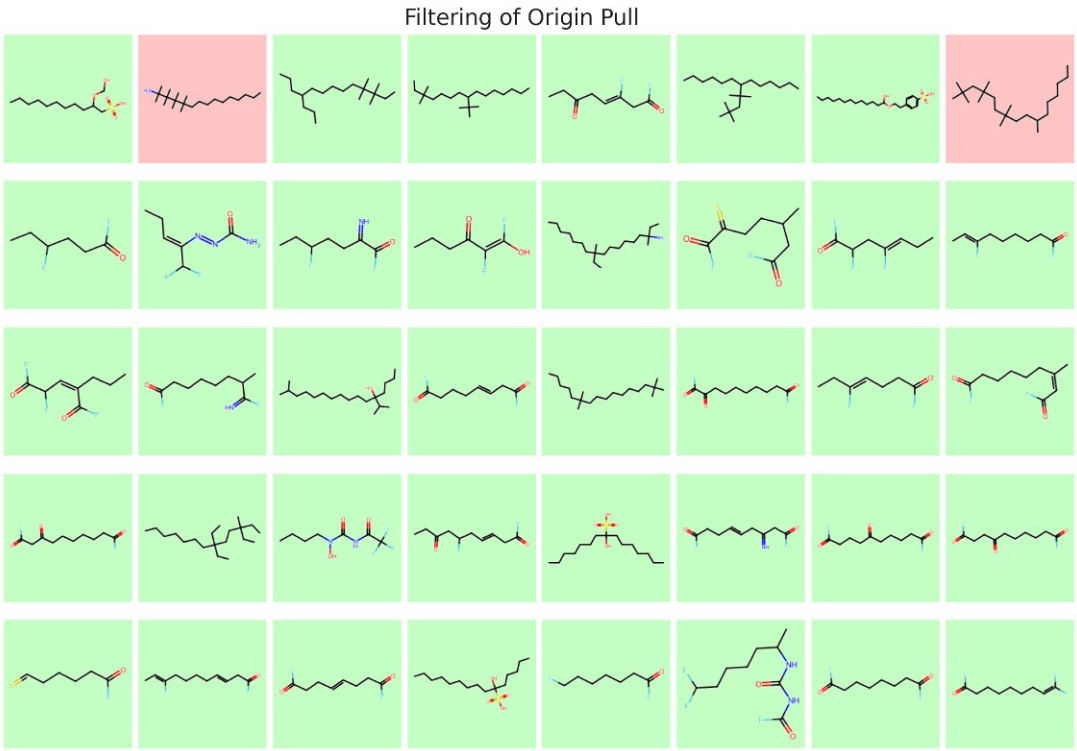

Figure B.8: Filter effect on Origin Pull

CHALMERS

UNIVERSITY OF TECHNOLOGY