A Computational Method to Support Chemical Product Design Based on Multi-objective Optimisation and Graph Transformers

Flavio S Correa da Silva[1*], Bilal Aslan[2*], Geoff Nitschke[2*]

[1]Department of Computer Science, University of Sao Paulo, Brazil
[2]Department of Computer Science, University of Cape Town, South Africa
*email: fcs@usp.br, aslbil001@myuct.ac.za, gnitschke@cs.uct.ac.za

Abstract

Chemical product design refers to the practice of developing novel chemical products given properties to be optimised and constraints to be satisfied. Strategies for chemical product design can be based on multi-objective constrained optimisation in a large search space of compounds whose properties are uncertain and partially known. Advances in machine learning, multi-objective optimisation, formal representation of chemical compounds and identified correlations between molecular structures and relevant properties, have fostered increased interest in computer-based techniques to identify candidate compounds for innovation in chemical products. In this paper we empirically explore a combination of state-of-the-art machine learning and evolutionary multi-objective optimisation methods to support chemical product design. In order to ground our arguments as concrete examples, we consider the design of domestic detergents, and explore how automating computational design can be controlled via specification of hyper-parameters, so as to generate solutions (detergents) with desired features. Our results contribute to the methodological problem of automating chemical product design, and more broadly functional molecular design.

Introduction

Chemical product design based on computational identification and optimisation of compounds given expected chemical properties has reduced design iterations and cycle times in comparison to trial-and-error synthesis methods (Chen and et al., 2018; Schneider, 2018). Computational identification and optimisation of compounds operates via iterative selection and modification of compounds to optimise selected properties of chemicals such as solvents, ionic liquids, polymers and medications (Ng and Gani, 2019). Desired properties can include, for example, low aquatic toxicity and favourable synthetic accessibility (Lysenko and et al., 2018; Zhuang and Ibrahim, 2021).

Traversal of the chemical design space to identify compounds of interest is computationally intractable, given that the chemical design space is estimated to contain over 10^{200} organic compounds (Reymond, 2015). Computational tractability has been managed by combining computational chemistry and machine learning, specifically, heuristics to optimise search space exploration (Keith and et al., 2021). This approach has been demonstrated to be effective in *de novo* molecular design and insight generation for drug discovery, materials science and pharmaceuticals (Bender and Cortes-Ciriano, 2021; Paul, 2021; Tkatchenko, 2020).

Various deep learning methods have been successfully applied for synthetic inference of properties and generation of compounds (Gawehn et al., 2016): auto-encoders have been trained to convert latent spaces as compound descriptor formats such as the *Simplified Molecular Input Line Entry Specification* (*SMILES*) (Gómez-Bombarelli and et al., 2018); graphs have been used to encode molecular structures (Samanta and et al., 2019); deep recurrent neural networks have been trained to generate chemically feasible innovative materials given molecular data (Yuan and et al., 2020) and to predict translation between reactants and products (Yuan and et al., 2020); and graph based generative (Gebauer et al., 2019) as well as Bayesian methods (Ikebata and et al., 2017) have enabled innovative organic compound synthesis. Many such computational product design methods are only partially automated, using various chemical data sets and molecular simulations to predict material properties of selected candidates (Curtarolo and et al., 2013; Pyzer-Knapp and et al., 2015). Prediction success rates are limited by the quality of data sets and require human expertise (Gómez-Bombarelli and et al., 2018).

Previous work has also demonstrated that evolutionary algorithms yield competitive results for chemical product design (Jensen, 2019; Kwon and et al., 2021; Leguy and al., 2009; Varela and Santos, 2022; Yoshikawa and et al., 2018), provided that the chemical design space is defined via specific molecular encoding. Selection and mutation are defined based on molecular fragments rather than atoms (Polishchuk, 2020) and specified using either grammars (Yoshikawa and et al., 2018; Nigam and et al., 2020) or statistical relations (Jensen, 2019).

The combination of machine learning to predict and evolutionary algorithms to optimise property values increases the likelihood of generation of synthetically viable compounds by accelerating stochastic search through evolutionary methods and improving property assessment through machine learning (Brown and et al., 2004; Le and Winkler, 2016), however ascertaining the most suitable combination of these approaches remains an open problem.

The development of new methods for automated chemical product design focusing on environmental sustainability is an issue of increasing importance. Thus, we have experimented with a method based on deep learning and evolutionary multi-objective optimisation (Belure et al., 2017; Winter and et al., 2019) to mitigate the effects of disposal of chemical products in the environment. Specifically, we have experimented with *Geometric Deep Learning* (Bronstein and et al., 2021) to estimate molecular properties and *Information-geometric Evolutionary Multi-objective Optimisation* (Ollivier and et al., 2017), to search through a chemical design space for sets of compounds that optimise multiple concurrent objectives such as minimisation of aquatic toxicity and maximisation of synthetic accessibility.

Innovations in the method introduced here are concentrated on the procedures to select the initial set of candidate compounds – which determine the region in the chemical design space to be explored to identify optimised compounds – and to define the seed compounds based on which the following generations of solutions are built. These procedures have been designed to ensure that the chemical design space is sufficiently explored (searched) to identify good candidate solutions.

In this study, we empirically evaluate our proposed method using a specific chemical product design case study – namely, the development of novel *detergents for domestic use* – and build sets of compounds that minimise indicators of toxicity and maximise indicators of synthetic accessibility, given an initial compound with the functionality of a known detergent. Results present the impact of chemical compound design experiments using various hyper-parameter settings to elucidate suitable method configurations for the given problem space.

In section II (Methods to Optimise Product Design), we review the proposed computational method, based on *Geometric Deep Learning* and *Information-geometric Optimisation*. In section III (Optimised Design of Detergents), we describe the problem of development of novel detergents, how it can be solved using the proposed method, and our obtained empirical results. In section IV (Discussion), we present a brief discussion. Section V (Conclusions and Future Work), presents conclusions and proposed future work.

II Methods to Optimise Product Design

A chemical design space is a set of compounds whose properties have been estimated using various techniques. It can be assumed that available properties are reliable, however it can also be expected that some values are missing. Chemical design spaces feature interesting structuring that can be explored to search for compounds that have specific properties. Specifically, it has been empirically validated that molecular properties correlate with particular patterns in their 3D structures, and consequently molecules featuring similar patterns also have similar properties (Crum-Brown and Fraser, 1865), an assumption known as the *Similar Property Principle* (Mitchell, 2014).

As a consequence of the *Similar Property Principle*, once appropriate patterns are identified for a given property, they can be used to characterise a distance relation between molecules, based on which properties can be estimated for all molecules in a chemical design space, based on the additional assumption that unknown properties of a molecule can be recurrently inferred given the (known) properties of their neighbours.

Once properties are estimated for all molecules in a design space, optimisation techniques can be employed to traverse the data set to find the most suitable compounds given properties to be optimised. What characterises a compound as a good product (for a given task) can be a complex combination of several properties, and user requirements typically focus on a relatively small subset of properties to be optimised. Hence, good heuristics to generate novel products start with a set $\mathcal{M}_0$ of known compounds that work reliably for the given task, and then search for alternative compounds to optimise the user specified properties within a region of the design space such that the distance between every molecule in the region and at least one $M_0 \in \mathcal{M}_0$ is bounded by a value $\hat{T}_0$, thus ensuring that the general properties of M_0 are preserved up to an acceptable approximation.

The two issues we address to build solutions for such heuristics are: (1) how to identify the appropriate patterns to define distances between molecules given properties of interest, and then to infer missing properties based on observed patterns, and (2) how to traverse the design space starting from $\mathcal{M}_0$ in such way that all properties of interest are optimised simultaneously and distances between solutions and at least one $M_0 \in \mathcal{M}_0$ are bounded by $\hat{T}_0$.

Our proposed heuristics is based on the *Tanimoto similarity* (Bajusz et al., 2015) between *molecular fingerprints* (Cereto-Massagué and et al., 2015). A molecular fingerprint is based on *features vectors*, listing selected substructures and connections between substructures such that a specific compound can be characterised as a Boolean vector (indicating presence or absence of each feature in the compound). The Tanimoto similarity between compounds M_i and M_j is defined as $T_{i,j} = \frac{k}{i+j+k}$ in which i is the number of features present in the fingerprint of M_i, where j is the number of features present in the fingerprint of M_j, and k is the number of features present in both fingerprints.

Tanimoto similarity is available from open access data sets such as *PubChem* (Kim and et al., 2016), which contains over 10^8 compounds. *PubChem* also contains *workhorse* property estimates for all compounds, used as coarse approximations for properties of interest, for example, toxicity can be estimated using *XLogP* (high *XLogP* suggests low toxicity) and synthetic accessibility can be estimated using *molecular complexity* and *molecular weight* (low complexity and weight suggest high accessibility).

Fine grained property estimates require more refined models, for example, *Geometric Deep Learning* (Bronstein and et al., 2021), where problem representation and solutions are co-optimised and a model for identification of patterns and inference of properties can be built to estimate properties for unforeseen molecules with improved accuracy. The problem with this approach is that it requires large data sets to build a model, as well as extensive computational resources.

Given the prohibitive computational costs of models using *PubChem*[1], we build accurate estimates for toxicity, using a molecular data set (containing 251k molecules) provided by a private company[2]. A small percentage (2%) of the data set was previously classified with respect to aquatic toxicity, based on which a Boolean decision model classifying molecules as *toxic* or *non-toxic* was built. At least two molecules shared the same *SMILES* description, despite being distinct molecules, confirming that Tanimoto similarity based on fingerprints extracted from *SMILES* representation can be too coarse if accurate property estimates are required. Namely, *2-ethoxy-N-hydroxybenzamidine* and *2-ethoxy-N'-hydroxybenzenecarboximidamide* have the same *SMILES* descriptor (`CCOC1=CC=CC=C1C(=NO)N`), but the former is classified as *non-toxic* and the latter is classified as *toxic*.

Our method uses Tanimoto similarity and the similarity obtained via *Geometric Deep Learning*: given an initial set of compounds $\mathcal{M}_0$ and a distance $\hat{T}_0$ that specifies a similarity measure threshold, Tanimoto similarity is employed to retrieve from *PubChem* the set of compounds M_i such that Tanimoto similarity between at least one $M_0 \in \mathcal{M}_0$ and M_i obeys the following inequality: $T_{M_0,M_i} \leq \hat{T}_0$. From this set of compounds, fine-grained similarity measures obtained via *Geometric Deep Learning* are used to rule out compounds identified as toxic. The final set characterises the search space from which the subset of optimal compounds with respect to user specified properties is identified. This uses an *Information-geometric Evolutionary Multi-objective Optimisation* algorithm as described in the following.

Recent machine learning advances have led to the development of *transformers*, achieving impressive results in natural language processing, machine translation, and image analysis (Zhang and et al., 2021). Transformers utilise *Self-Attention Mechanisms* (*SAM*) to explore data organised in simple structures, such as sequence relations in texts and neighbourhood relations in image segments, and build more general, goal-oriented relations. More sophisticated data organisation, such as what can be represented using graphs (Bronstein and et al., 2021), has proven to be challenging for transformers. Graph transformers have been developed to address this based on purpose-oriented forms of graph encoding, leading to advancements in molecular property prediction by incorporating topological and geometric information representing molecular structure, based on the *Similar Property Principle* (Mitchell, 2014).

Recent approaches directly integrate graph structural information into transformers via implementing improved positional encoding and improved attention maps derived from graph topology (Cai and Lam, 2020; Hussain et al., 2021; Mialon and et al., 2021; Ying and et al., 2021), and exploration of 3D molecular structure, specifically considering inter-atomic distance as geometric information to be explored in attention maps (Zhou and et al., 2023).

Our method extends *Uni-Mol* (Zhou and et al., 2023), with superior empirical accuracy in comparison with competing architectures on benchmark data sets for molecular property prediction. The *Uni-Mol* method uses a transformer based backbone with incorporated invariant spatial positional encoding and pair-level representation to effectively capture 3D information. Unlike previous molecular pre-training models (Hu and et al., 2019; Li and et al., 2021; Wang and et al., 2022), *Uni-Mol* treats a molecule as a set of 3D nodes with atom type and 3D coordinates, rather than a graph. Previous models used a spatial local-connected graph to represent 3D nodes (Schütt and et al., 2017; Gasteiger et al., 2020, 2021; Liu and al., 2022), which may not capture effectively long-range atomic interactions. *Uni-Mol* leverages transformers to capture such interactions using a Pre-LayerNorm architecture to handle 3D spatial data (Xiong and et al., 2020), invariant spatial positional encoding, pair representation, and an SE(3)-Equivariance coordinate head.

[1] https://pubchem.ncbi.nlm.nih.gov/

[2] Smarter Sorting Inc:https://www.smarterx.com/

Positions in three-dimensional space are real valued, hence positional encoding has to be invariant under global rotation and translation. To achieve this, *Uni-Mol* adapts relative positional encoding by using Euclidean distances between atom pairs, and a pair type aware Gaussian kernel (Shuaibi and et al., 2021). The *D*-channel positional encoding of atom pair ij is denoted as $p_{ij} = \{G(A(d_{ij}, t_{ij}; \mathbf{a}, \mathbf{b}), \mu^k, \sigma^k) | k \in [1, D]\}, A(d, r; \mathbf{a}, \mathbf{b}) = a_r d + b_r$ where the Gaussian density function is $G(d, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-[(d-\mu)^2/(2\sigma^2)]}$, μ and σ are Gaussian density function parameters, d_{ij} is the Euclidean distance of atom pair ij, and t_{ij} is the pair-type of atom pair ij. The pair-type here is not the chemical bond, which is instead determined by atom types of pair ij. $A(d_{ij}, t_{ij}; \mathbf{a}, \mathbf{b})$ is the affine transformation with parameters **a** and **b**, d_{ij} corresponding to its pair-type t_{ij}.

Transformers are designed to maintain both token-level and pair-level representations. Token-level representation is used as a baseline for fine-tuning in downstream tasks. However, the spatial positions of the input are encoded at the pair-level, which enables the model to better capture the 3D spatial relationships between atoms. To initialise the pair-level representation, a spatial positional encoding is employed. The atom-to-pair communication is achieved through the use of multi-head *SAM* in the form of query-key products. This allows for updating of the pair-level representation and further refinement of characterisation of complex spatial relationships between atoms.

The update of ij pair representation is denoted as $q_{ij}^0 = p_{ij}M, q_{ij}^{l+1} = q_{ij}^l + \{\frac{Q_i^{l,h}(K_j^{l,h})^T}{\sqrt{d}} | h \in [1, H]\}$ where q_{ij}^l is the pair representation of atoms ij in l-th layer, H is the number of attention heads, d is the dimension of hidden representations, $Q_i^{l,h}(K_j^{l,h})$ is the Query-Key of the i-th (j-th) atom in the l-th layer h-th head, and $M \in RD \times H$ is the projection matrix to make the representation the same shape as multi-head Query-Key product results. *Uni-Mol* incorporates the 3D information into the atom representation through the use of pair-to-atom communication. This is achieved by utilising the pair representation as a bias term in *SAM*. This allows for the propagation of the pair-level information to the atom-level. *SAM* with pair-to-atom communication is denoted as $Attention(Q_i^{l,h}, K_j^{l,h}, V_j^{l,h}) = softmax(\frac{Q_i^{l,h}(K_j^{l,h})^T}{\sqrt{d}} + q_{ij}^{l-1,h})V_j^{l,h}$ where $V_j^{l,h}$ is the j-th atom in the l-th layer h-th head.

The use of 3D spatial positional encoding and pair representation in *Uni-Mol* enhances its ability to capture the complex spatial relationships between atoms.

However, the model still lacks the capability to directly output 3D coordinates, which is crucial for tasks that require 3D spatial information. To address this limitation, a SE(3)-Equivariance head is used in *Uni-Mol*. This addition allows for the direct output of 3D coordinates and enhances the model's overall performance in tasks that require 3D spatial information: $\hat{x}_i = x_i + \sum_{j=1}^{n} \frac{(x_i - x_j)c_{ij}}{n}, c_{ij} = ReLU((q_{ij}^L - q_{ij}^0)U)W$ where *n* is the number of total atoms, *L* is the number of layers in model, $x_i \in R^3$ is the input coordinate of *i*-th atom, and $\hat{x}_i \in R^3$ is the output coordinate of *i*-th atom, $ReLU(y) = max(0, y)$ is Rectified Linear Unit, $U \in R^{H \times H}$ and $W \in R^{H \times 1}$ are the projection matrices to convert pair representation to a scalar.

In order to effectively pre-train *Uni-Mol*, we used a large-scale data set of organic molecules, where molecular pre-training data consisted of approximately 19 million molecules, which were sourced from multiple public data sets. To obtain the 3D conformations, a combination of ETKGD (Riniker and Landrum, 2015) and Merck Molecular Force Field optimisation (Halgren, 1996) from RDKit tool (Landrum and et al., 2013) was used to randomly generate ten conformations for each molecule. Additionally, a 2D conformation was generated to address rare cases where 3D conformations could not be generated.

Self-supervised learning is crucial for effective learning from large-scale unlabeled data. *Uni-Mol* utilises a masked atom prediction task as its self-supervised objective. For each molecule or pocket, a special atom $[CLS]$ is added to represent the entire molecule, with its coordinate being the centre of all atoms. However, as 3D spatial positional encoding contains pair distances, the corresponding atom types could be inferred easily, and therefore, the masked atom prediction cannot encourage the model to learn useful information To overcome this limitation and encourage learning from 3D information, a 3D position denoising task was designed. This task involves adding uniform noise of $[-1\dot{A}, 1\dot{A}]$ to the coordinates of 15% of the randomly selected atoms, after which the spatial positional encoding is calculated based on the corrupted coordinates.

Two additional heads were also employed to recover the correct spatial positions. The first head, the pair-distance prediction head, uses the pair representation to predict the correct Euclidean distances of the atom pairs with corrupted coordinates. The second head, the coordinate prediction head, utilises the SE(3)-Equivariance coordinate head to predict the correct coordinates for the atoms with corrupted coordinates. The overall pre-training process and architecture of *Uni-Mol* are illustrated in figure 1.

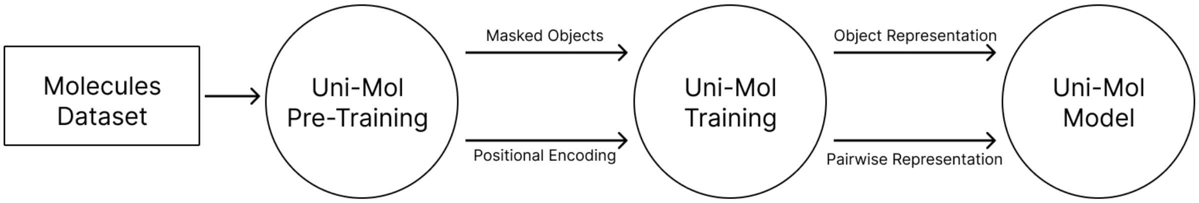

Figure 1: *Uni-Mol* graph transformer. Left: Pre-training architecture. Middle: Inputs, including masked objects and spatial positional encoding created by pairwise Euclidean distances are used for training. Right: Pairwise and individual object representations comprise foundations for model.

The interested reader is advised to check the original presentation of *Uni-Mol* for additional details (Zhou and et al., 2023). To maintain consistency with the pre-training process, the same data pre-processing pipeline was employed during fine-tuning. For molecules, multiple random conformations can be generated in a short time, making it possible to use them as data augmentation during fine-tuning to enhance performance and robustness. In cases where 3D conformations could not be generated, the molecular graph was used as a 2D conformation. Similar to natural language processing and image analysis, the representation of $[CLS]$, which represents the entire molecule or the mean representation of all atoms, was used in conjunction with a linear head to fine-tune on downstream tasks.

Once a set of candidate compounds which are sufficiently similar to the initial compounds in $\mathcal{M}_0$ is selected according to the criteria described in previous paragraphs and toxic molecules are removed from this set, the next step in our method is the identification of a *solution set* containing only optimal compounds. Broadly, we characterise our problem as an *Information-geometric Evolutionary Multi-objective Optimisation* task with imperfect information:

- **Given** a set $\mathcal{A}$ of relevant properties which can be ascribed to specified compounds (and which are assumed to have domains ranging through real-valued intervals); a subset $\mathcal{A}_{opt} \subseteq \mathcal{A}$ of those properties which must be *optimised*, i.e. either minimised or maximised; a subset $\mathcal{A}_{constr} \subseteq \mathcal{A}$ of those properties which define *constraints*, i.e. such that for each property $A_c \in \mathcal{A}_{constr}$ we have defined two values v_c^{min}, v_c^{max}, $v_c^{min} \leq v_c^{max}$; and a set of compounds to be considered as candidate solutions for the problem;
- **Find** a set of compounds which are *good enough approximations* of the compounds that optimise the properties in $\mathcal{A}_{opt}$ while ensuring that properties $A_c \in \mathcal{A}_{constr}$ belong to the interval $[v_c^{min}, v_c^{max}]$.

The strategy to navigate towards near-optimal compounds given specified properties $\mathcal{A}_{opt}$ and $\mathcal{A}_{constr}$ follows the conceptual framework of *Multi-objective Covariance Matrix Adaptation Evolution Strategy – Mo-CMA-ES* (Igel et al., 2007), adapted to a non-parametric setting, without mutations and with selection defined for whole molecules.

Starting from the set of candidate compounds, given a *generic threshold* $\hat{T} \geq \hat{T}_0$, we retrieve from the candidate compounds the subset $\mathcal{M}_0^c = \{M : \exists M_0 \in \mathcal{M}_0 : T_{M_0,M} \geq \hat{T}\}$. By definition, $\mathcal{M}_0 \subseteq \mathcal{M}_0^c$. For each $M_i \in \mathcal{M}_0^c$, we check whether the constraints defined for properties in $\mathcal{A}_{constr}$ are satisfied, and build $\tilde{\mathcal{M}}_0^c \subseteq \mathcal{M}_0^c$ containing only the compounds that satisfy all constraints. From these, we build the *Pareto frontier* of candidate solutions which comprise a Pareto equilibrium considering all properties in $\mathcal{A}_{opt}$ estimated according to "workhorse" values readily available from *PubChem*, this way building the initial solution set $\tilde{\mathcal{S}}_0 = \{M_i : M_i \in \textit{Pareto frontier}\}$. As a final step, we rule out from $\tilde{\mathcal{S}}_0$ those compounds identified as toxic using the estimates provided by the graph Transformers, thus building the *final Pareto frontier* $\mathcal{S}_0$.

Given a *generation size* λ, we select at random M_{01}, ..., $M_{0\lambda}$ from $\mathcal{S}_0$, to build the *offspring set of compounds* $\mathcal{M}_1$, which is used respectively to build $\mathcal{M}_1^c, \tilde{\mathcal{M}}_1^c, \tilde{\mathcal{S}}_1$ and $\mathcal{S}_1$. This procedure is repeated to build $\mathcal{S}_2, \mathcal{S}_3 \ldots$, until some stability criteria is reached in $\mathcal{S}_N$ for some finite N – for example, until $\frac{|\mathcal{S}_{k+1}|}{|\mathcal{S}_k|} \approx 1$. To help avoid local optima, we also include, following the strategy of *Mo-CMA-ES*, a *growth factor* $\hat{G} > 1$ for λ: if $\frac{|\mathcal{S}_{k+1}|}{|\mathcal{S}_k|} < 1$, then λ is updated to $\lambda \times \hat{G}$, and if $\frac{|\mathcal{S}_{k+1}|}{|\mathcal{S}_k|} > 1$, then λ is updated to $\frac{\lambda}{\hat{G}}$.

The final solution set $\mathcal{S}_N$ obtained comprises the compounds suggested as potential solutions upon discretion of a human product designer. Figure 2 presents an overview of this chemical design (discovery and optimisation) process.

III Optimised Design of Detergents

In order to work on a real-world problem, we have focused on the development of new *detergents for domestic use*. Detergents are built from compounds with peculiar molecular configurations, for which the *Similar Property Principle* is valid (Smulders and et al., 2002). Molecules that are used in detergents typically have a strip-like shape, sometimes

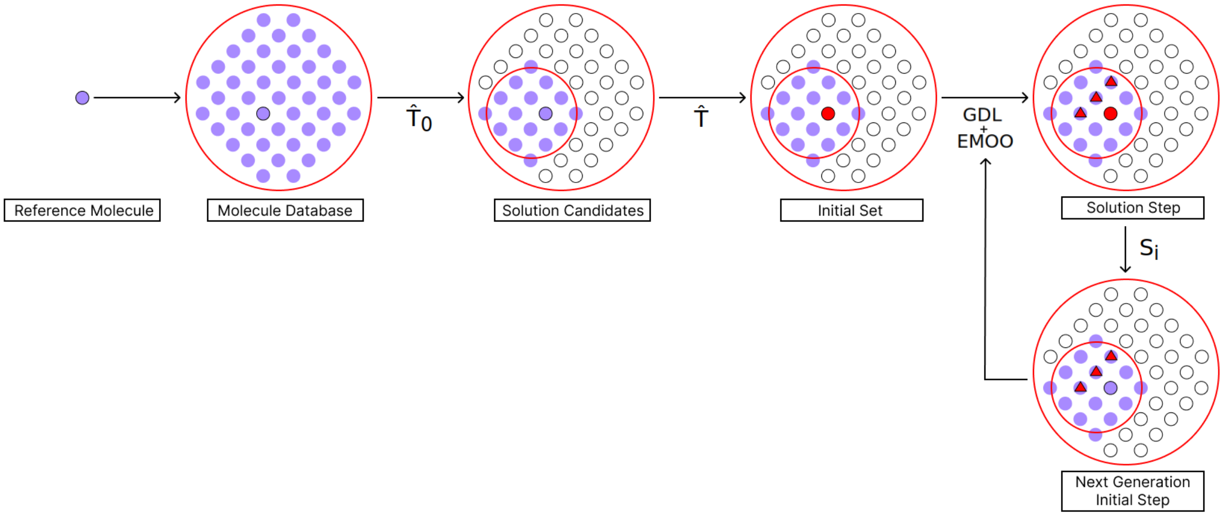

Figure 2: Overview of proposed method: given an initial chemical design space, a set of candidate solutions is selected based on Tanimoto similarity $\hat{T}_0$; from this set, initial compounds are identified based on Tanimoto similarity $\hat{T}$; using the initial compounds, high-risk compounds are removed using *Geometric Deep Learning* (GDL) and optimised compounds are identified using *Evolutionary Multi-objective Optimisation* (EMOO), thus building a solution set; the obtained solution set is used as a new set of initial compounds to iterate the process and build new generations of optimised compounds, until stability is reached; the final result is the set of suggested compounds for consideration for product design.

with a bifurcation. This forms either a V-like or an Y-like shape and features *lipophilic* (*hydrophobic*) behaviour at one end, or at the bifurcated end when a bifurcation occurs, and hydrophilic behaviour at the other. For example, when a liquid detergent is applied on a greasy surface (for example, a frying pan that has just been used for cooking) forming a film, the *lipophilic* (*hydrophobic*) end of molecules is attracted by the greasy surface and the hydrophilic end remains free and away from the surface. When the film is washed with running water, the hydrophilic end of detergent molecules is pulled away with the current, bringing together with it the grease and thus cleaning the surface.

In this section, *XLogP* measures the ratio between *lipophilicity* and *hydrophilicity* of a compound, and can be maximised to build an approximate indication of low toxicity and good cleaning properties. Toxicity estimates can be further refined using graph transformers, as outlined in previous sections. Considering that a chemical product must be manufactured via industry, it is important to assess manufacturing costs. Typically, these costs are inversely correlated with *synthetic accessibility*, which denotes the required effort to synthesise a compound. Synthetic accessibility, in turn, is approximately inversely correlated with *Molecular Weight* and with *Molecular (structural) Complexity*. Hence, manufacturing costs can be approximately assessed based on these two properties.

The properties to be optimised can be, therefore, *XLogP* (to be maximised), and *Complexity* and *Molecular Weight* (to be minimised). Additionally, we consider a Boolean constraint provided by graph Transformers, that classifies compounds as either toxic or non-toxic. In our experiments, we simulate the initial set of compounds $\mathcal{M}_0$ by starting with a specific compound M_{init} which has been used previously in a patented detergent, namely *Methylhexadecyl hydrogen sulphate*, whose *SMILES* representation is given by `CCCCCCC(C)CCCCCCCCCOS(=O)(=O)O` and which is present in a detergent mixture that has been patented in 2015 (Ellison and et al., 2015), and then selecting ten compounds at random from *PubChem* featuring Tanimoto similarity at least 97% with respect to M_{init}.

There are three hyper-parameters that control the behaviour of the traversal of the set of candidate compounds:

1. Initial similarity threshold $\hat{T}_0$: larger $\hat{T}_0$ ensure that solutions will be similar to M_0, which can be good when priority is given to preserving the properties of M_0 in final solutions, at the cost of reducing possibilities for optimisation of properties of interest.

2. Generic similarity threshold $\hat{T}$: larger $\hat{T}$ slow down convergence, possibly inducing the method to go through additional cycles in order to reach stabilisation.

3. Generation size λ: larger λ decrease the randomness in selection of compounds to act as initial compounds in next

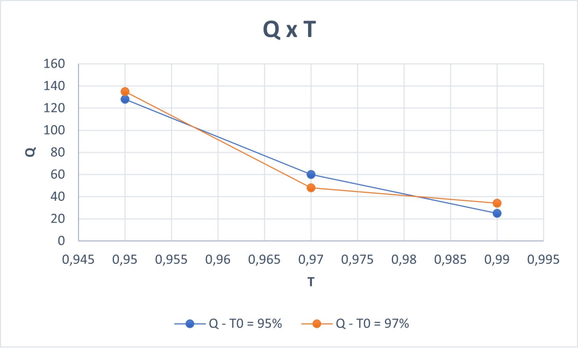

Figure 3: Cardinality of solution sets according to $\hat{T}$ and $\hat{T}_0$

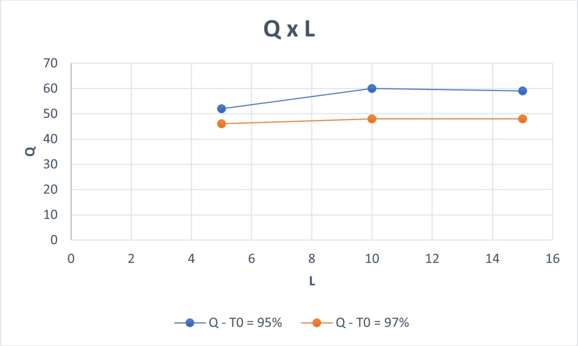

Figure 4: Cardinality of solution sets according to λ and $\hat{T}_0$

generations – in the limit, if $\lambda \geq$ the cardinality of the present solution set, randomness is eliminated completely.

We experimented with various values for these hyper-parameters, specifically: $\hat{T}_0 \in [0.95, 0.97], \hat{T} \in [0.95, 0.99]$ and $\lambda \in [5, 15]$. As expected, smaller $\hat{T}_0$ and of $\hat{T}$ generated more comprehensive solution sets with remarkably few exceptions (less than 2% of the compounds in each solution set), solution sets obtained with larger $\hat{T}$ were subsets of solution sets obtained with smaller $\hat{T}$.

Figure 3 presents the cardinality of final solution sets, for $\hat{T} \in \{0.95, 0.97, 0.99\}$ and $\hat{T}_0 \in \{0.95, 0.97\}$. In these experiments, we have adopted $\lambda = 10$. Where, different λ influence the selection of compounds in intermediate generations during traversal of the set of candidate compounds, and do not have direct influence of comprehensiveness of solution sets. Figure 4 presents the cardinality of final solution sets, for $\lambda \in \{5, 10, 15\}$ and $\hat{T}_0 \in \{0.95, 0.97\}$. In these experiments, we have adopted $\hat{T} = 0.97$.

As a qualitative metric, we check what compounds are in the final solution sets, observing that 14 compounds were present in all solution sets, irrespective of choice of hyper-parameter values. Five of these compounds belong to patented detergents, namely:

1. - *2-Methylpentadecyl hydrogen sulfate*
   - (`CCCCCCCCCCCCCC(C)COS(=O)(=O)O`)
   - patent issued in 2002 (Kvietok and et al., 2002),

2. - *20-Methyldocosyl hydrogen sulfate*
   - (`CCC(C)CCCCCCCCCCCCCCCCCCCOS(=O)(=O)O`)
   - patent issued in 2014 (Scheibel and et al., 2014),

3. - *2-Methylhexadecyl hydrogen sulfate*
   - (`CCCCCCCCCCCCCCC(C)COS(=O)(=O)O`)
   - patent issued in 2015 (Federle and et al., 2015),

4. - *2-Octylundecyl hydrogen sulfate*
   - (`CCCCCCCCCC(CCCCCCCC)COS(=O)(=O)O`)
   - patent issued in 2019 (Holland and et al., 2019),

5. - *2-Octyldecyl hydrogen sulfate*
   - (`CCCCCCCCC(CCCCCCCC)COS(=O)(=O)O`)
   - patent issued in 2020 (Holland et al., 2020).

All other compounds, despite not having identified previous patents in *PubChem* indicating their use as detergents, share structural similarities with each other.

IV Discussion

This study investigated the optimisation of a specific chemical product, as a specific case study representative of the larger field of automated chemical product design. That is, given a space of candidate solutions for a specific problem organised as a graph in which edge labels denote similarities between pairs of candidates, and given a collection of properties to be optimised and reference points in the space of solutions, we find solution sets comprising points within a region defined by balls centered on each of the reference points. Solution sets are defined in such way that there are no two points M_i and M_j in a solution set where M_i is better than M_j at every property being optimised.

In this case study of chemical product design, we include the availability of more than one available procedure to estimate and verify properties of candidate solutions, considering that accurate procedures can be highly expensive and inexpensive procedures can be inaccurate. Specifically, in our method, candidate solutions are compounds found in a large scale data set such as *PubChem*, and the reference point is given by a set of compounds which are active elements in previously known (patented) detergents. This ensures that all solutions are sufficiently similar to at least one detergent belonging to the reference set and will thus share properties with it. This ensures all such solutions are worth checking as potential candidates for novel detergents (with optimal properties), where properties to be optimised are acceptable proxies for attributes found in high quality real-world products such as low risk of aquatic toxicity and low manufacturing costs.

This discovery of existing detergents, with optimised properties (section III), was enabled by our extension of *Uni-Mol* coupled with *Information-geometric Evolutionary Multi-objective Optimisation* (section II). Specifically, we applied the Tanimoto similarity measure between pairs of compounds to navigate across a space of candidate solutions and to subsequently build precise estimates for risks related to aquatic toxicity, based on *Geometric Deep Learning*, which are then used to eliminate (from solution sets), those compounds with high toxicity risk. Our empirical results indicate that this method is capable of identifying potentially risky compounds which are not noticed using less accurate methods. To effectively and efficiently manage exploration of candidate solution space, our method can be fine tuned using various hyper-parameters (section II).

The effectiveness of the method presented here was demonstrated via the identification of chemical compound solutions which had been previously identified using standard chemical synthesis processes (Chen and et al., 2018; Schneider, 2018). This suggests that our method is suitable for discovering novel solutions that are structurally similar to previously identified solutions (for example, current detergents synthesized by chemical product design), but with further optimised properties (such as lower toxicity). Figure 2 presents an overview of this chemical product design process, where the optimised compounds discovered in our solution set were equivalent to 14 existing compounds (already patented detergents highlighted in section 3: *Optimised Design of Detergents*). This result further validates the efficacy of our method as an assistive computational chemical design tool and as a molecular property optimisation tool.

Overall, results obtained thus far indicate the potential of our method as a computational tool to assist with optimising chemical product designers, via rapid identification of novel solutions (compounds) with advantageous design properties. Current research is further validating our computational method on a broader range of chemical compound design and optimisation tasks. For example, comprehensive verification of some compounds belonging to the obtained solution sets which, however, have not been part of previous patents identified in *PubChem*, indicate that these compounds can cause skin and eye irritation, hence being possibly inappropriate as compounds to be included in domestic detergents. These attributes, however, have not been considered here, and will be considered in future experiments.

V Conclusions and Future Work

This study presented an empirical validation of a computational chemical design and optimisation method to assist with novel chemical product design. Our preliminary results indicate that our proposed method is particularly useful for identifying alternative optimal chemical design solutions given a known (near optimal) solution as a starting point. This study's experiments focused on the optimisation of molecular properties relevant to ensuring low environmental impact of product and production costs. The proposed computational method is sufficiently general to be applicable to solving various chemical design and optimisation (and more broadly molecular design) problems, given that all such design problems operate in solution spaces characterised by specific structural (molecular) definitions.

Future work will validate our method as a computational tool for assisting optimal chemical product design for various real-world applications, as a step towards automated molecular design. Specifically, we shall focus on using generative and predictive machine learning in combination with generative computational chemistry to devise novel molecular compounds to be used in pharmaceuticals or other functional chemical agents. We shall also focus on further validation of our method using quantitative analysis and comparative assessment with respect to existing approaches that could be used for the same design problems we have taken into consideration.

Acknowledgements

This research has been partially funded by the *Algorand Foundation*, *SmarterX* and the South African *National Research Foundation*.

References

Bajusz, D., Rácz, A., and Héberger, K. (2015). Why is Tanimoto Index an Appropriate Choice for Fingerprint-based Similarity Calculations? *Journal of cheminformatics*, 7(1):1–13.

Belure, S., Shir, O., and Vikas, N. (2017). Protein Design by Multi-objective Optimization: Evolutionary and Non-Evolutionary Approaches. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1081–1088, Berlin, Germany. ACM.

Bender, A. and Cortes-Ciriano, I. (2021). Artificial Intelligence in Drug Discovery: What Is Realistic, What Are Illusions? *Drug Discovery Today*, 26(1):1040.

Bronstein, M. and et al. (2021). Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv preprint arXiv:2104.13478*.

Brown, N. and et al. (2004). A Graph-based Genetic Algorithm and its Application to the Multiobjective Evolution of Median Molecules. *Journal of Chemical Information and Modeling*, 44(1):1079–1087.

Cai, D. and Lam, W. (2020). Graph Transformer for Graph-to-sequence Learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7464–7471.

Cereto-Massagué, A. and et al. (2015). Molecular Fingerprint Similarity Search in Virtual Screening. *Methods*, 71:58–63.

Chen, R. and et al. (2018). Machine Learning for Drug-Target Interaction Prediction. *Molecules*, 23(9):2208.

Crum-Brown, A. and Fraser, T. (1865). The Connection of Chemical Constitution and Physiological Action. *Transactions of the Royal Society of Edinburgh*, 25(1968-1969):257.

Curtarolo, S. and et al. (2013). The High-throughput Highway to Computational Materials Design. *Nature Materials*, 12(1):191–201.

Ellison, R. and et al. (2015). Laundry Detergent Composition and Method of Making Thereof. US Patent App. 14/402,327.

Federle, T. and et al. (2015). Intermediates and Surfactants useful in Household Cleaning and Personal Care Compositions, and Methods of Making the Same. US Patent 8,933,131.

Gasteiger, J., Becker, F., and Günnemann, S. (2021). GEMNET: Universal Directional Graph Neural Networks for Molecules. *Advances in Neural Information Processing Systems*, 34:6790–6802.

Gasteiger, J., Groß, J., and Günnemann, S. (2020). Directional Message Passing for Molecular Graphs. *arXiv preprint arXiv:2003.03123*.

Gawehn, E., Hiss, J., and Schneider, G. (2016). Deep Learning in Drug Discovery. *Molecular Informatics*, 35(1):3–14.

Gebauer, N., Gastegger, M., and Schütt, K. (2019). Symmetry-adapted Generation of 3D Point Sets for the Targeted Discovery of Molecules. In *Proceedings of the Conference on Neural Information Processing Systems*, Vancouver, Canada. ACM.

Gómez-Bombarelli, R. and et al. (2018). Automatic Chemical Design using a Data-driven Continuous Representation of Molecules. *ACS Central Science*, 4(2):268–276.

Halgren, T. (1996). Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *Journal of Computational Chemistry*, 17(5-6):490–519.

Holland, B., Bernhardt, R., and Sajic, B. (2020). Cold-water Laundry Detergents. US Patent 10,570,352.

Holland, B. and et al. (2019). Detergents for Cold-water Cleaning. US Patent 10,421,930.

Hu, W. and et al. (2019). Strategies for Pre-training Graph Neural Networks. *arXiv preprint arXiv:1905.12265*.

Hussain, M., Zaki, M., and Subramanian, D. (2021). Edge-augmented graph transformers: Global self-attention is enough for graphs. *arXiv preprint arXiv:2108.03348*.

Igel, C., Hansen, N., and Roth, S. (2007). Covariance Matrix Adaptation for Multi-Objective Optimization. *Evolutionary computation*, 15(1):1–28.

Ikebata, H. and et al. (2017). Bayesian Molecular Design with a Chemical Language Model. *Journal of Computer-Aided Molecular Design*, 31(4):379–391.

Jensen, J. (2019). A Graph-based Genetic Algorithm and Generative Model/Monte Carlo Tree Search for the Exploration of Chemical Space. *Chemical Science*, 10(12):3567–3572.

Keith, J. and et al. (2021). Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. *Chemical Reviews*, 121:9816–9872.

Kim, S. and et al. (2016). PubChem Substance and Compound Databases. *Nucleic acids research*, 44(D1):1202–1213.

Kvietok, F. and et al. (2002). Detergent Composition Comprising Mid-chain Branched Surfactants. US Patent 6,482,789.

Kwon, Y. and et al. (2021). Evolutionary Design of Molecules based on Deep Learning and a Genetic Algorithm. *Nature Scientific Reports*, 11(17304):4–6.

Landrum, G. and et al. (2013). RDKit: A Software Suite for Cheminformatics, Computational Chemistry, and Predictive Modeling. *https://www.rdkit.org/*.

Le, T. and Winkler, D. (2016). Discovery and Optimization of Materials using Evolutionary Approaches. *Chemical Reviews*, 116(1):6107–6132.

Leguy, J. and al. (2009). EVOMOL: A Flexible and Interpretable Evolutionary Algorithm for Unbiased de novo Molecular Generation. *Journal of Cheminformatics*, 12(55):1–19.

Li, P. and et al. (2021). An Effective Self-supervised Framework for Learning Expressive Molecular Global Representations to Drug Discovery. *Briefings in Bioinformatics*, 22(6):bbab109.

Liu, Y. and al. (2022). Spherical Message Passing for 3D Molecular Graphs. In *International Conference on Learning Representations (ICLR)*.

Lysenko, A. and et al. (2018). An Integrative Machine Learning Approach for Prediction of Toxicity-related Drug Safety. *Life Science Alliance*, 1(6):e201800098.

Mialon, G. and et al. (2021). Graphit: Encoding Graph Structure in Transformers. *arXiv preprint arXiv:2106.05667*.

Mitchell, J. (2014). Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(5):468–481.

Ng, K. and Gani, R. (2019). Chemical Product Design: Advances in and Proposed Directions for Research and Teaching. *Computers & Chemical Engineering*, 126:147–156.

Nigam, A. and et al. (2020). Augmenting Genetic Algorithms with Deep Neural Networks for Exploring the Chemical Space. In *Proceedings of the Eighth International Conference on Learning Representations*, Addis Ababa, Ethiopia. ACM.

Ollivier, Y. and et al. (2017). Information-geometric Optimization Algorithms: A Unifying Picture via Invariance Principles. *Journal of Machine Learning Research*, 18(18):1–65.

Paul, D. (2021). Artificial Intelligence in Drug Discovery and Development. *Drug Discovery Today*, 26(1):80–93.

Polishchuk, P. (2020). CReM: Chemically Reasonable Mutations Framework for Structure Generation. *Journal of Cheminformatics*, 12(1):28.

Pyzer-Knapp, E. and et al. (2015). What is High-throughput Virtual screening? A Perspective from Organic Materials Discovery. *Annual Review of Materials Research*, 45(1):195–216.

Reymond, J. (2015). The Chemical Space Project. *Accounts of Chemical Research*, 48:722–730.

Riniker, S. and Landrum, G. (2015). Better Informed Distance Geometry: Using What we Know to Improve Conformation Generation. *Journal of Chemical Information and Modeling*, 55(12):2562–2574.

Samanta, B. and et al. (2019). NeVAE: A Deep Generative Model for Molecular Graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1110–1117, Honolulu, USA. ACM.

Scheibel, J. and et al. (2014). Compositions comprising a near terminal-branched compound and methods of making the same. US Patent 8,883,698.

Schneider, G. (2018). Automating Drug Discovery. *Nature Reviews Drug Discovery*, 17(2):97–113.

Schütt, K. and et al. (2017). SchNet: A Continuous-filter Convolutional Neural Network for Modeling Quantum Interactions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 992–1002, Long Beach, USA. ACM.

Shuaibi, M. and et al. (2021). Rotation Invariant Graph Neural Networks using Spin Convolutions. *arXiv preprint arXiv:2106.09575*.

Smulders, E. and et al. (2002). *Laundry Detergents*. Wiley Online Library.

Tkatchenko, A. (2020). Machine Learning for Chemical Discovery. *Nature Communications*, 11:4125.

Varela, D. and Santos, J. (2022). Niching Methods Integrated with a Differential Evolution Memetic Algorithm for Protein Structure Prediction. *Swarm and Evolutionary Computation*, 71(1):101062.

Wang, Y. and et al. (2022). Molecular Contrastive Learning of Representations via Graph Neural Networks. *Nature Machine Intelligence*, 4(3):279–287.

Winter, R. and et al. (2019). Efficient Multi-objective Molecular Optimization in a Continuous Latent Space. *Chemical Science*, 10(34):8016–8024.

Xiong, R. and et al. (2020). On Layer Normalization in the Transformer Architecture. In *International Conference on Machine Learning*, pages 10524–10533. PMLR.

Ying, C. and et al. (2021). Do Transformers Really Perform Badly for Graph Representation? *Advances in Neural Information Processing Systems*, 34:28877–28888.

Yoshikawa, N. and et al. (2018). Population-based de novo Molecule Generation using Grammatical Evolution. *Chemistry Letters*, 47(11):1431–1434.

Yuan, Q. and et al. (2020). Molecular Generation Targeting Desired Electronic Properties via Deep Generative Models. *Nanoscale*, 12(12):6744–6758.

Zhang, A. and et al. (2021). Dive into Deep Learning. *arXiv preprint arXiv:2106.11342*.

Zhou, G. and et al. (2023). Uni-Mol: A Universal 3D Molecular Representation Learning Framework. *Eleventh International Conference on Learning Representations (ICLR 2023)*.

Zhuang, D. and Ibrahim, A. (2021). Deep Learning for Drug discovery: A study of Identifying High Efficacy Drug Compounds using a Cascade Transfer Learning Approach. *Applied Sciences*, 11(1):7772.