Multi-objective Evolution for Automated Chemistry

Bilal Aslan
Department of Computer Science
University of Cape Town
Cape Town, South Africa
aslbil001@myuct.ac.za

Flavio S Correa da Silva
Department of Computer Science
University of Sao Paulo
Sao Paulo, Brazil
fcs@usp.br

Geoff Nitschke
Department of Computer Science
University of Cape Town
Cape Town, South Africa
gnitschke@cs.uct.ac.za

***Abstract*—A fundamental problem in chemical product design is how to suitably identify chemical compounds that optimise multiple properties for a given application whilst satisfying relevant constraints. Current product synthesis generally uses trial-and-error experimentation, requiring lengthy and expensive research and development efforts. This paper introduces a novel computational chemistry approach for product design combining geometric deep learning for inference of property values and evolutionary multi-objective optimisation for identification of products of interest. Preliminary empirical results indicate that the proposed approach can be used to optimise product design considering multiple objectives and constraints given incomplete molecular attribute information.**

***Index Terms*—Evolutionary Multi-Objective Optimisation, Computational Chemistry, Geometric Deep Learning**

I. Introduction

Novel product design based on computational molecular discovery [1] has demonstrated significantly reduced design iterations, shorter iteration cycle times and increased chemical synthesis in comparison to trial-and-error chemical synthesis methods [2]. Computational molecular design achieves chemical synthesis via iterative selection and modification of compounds to optimise desired attributes of products such as solvents, ionic liquids, polymers and medications [3]. For example, target attributes in pharmacokinetics can include low toxicity and favourable synthetic accessibility [4].

Deep Learning [5] (DL) has been successfully applied for synthetic compound generation and attribute value inference [6], [7]. For example, auto-encoders have been trained to convert latent-spaces to molecule descriptors (such as SMILES: *Simplified Molecular Input Line Entry Specification* [8]) and encode molecular solutions [9]. Recent DL advances have resulted in Transformer-based methods, with impressive results in natural language processing, machine translation, and image analysis [10]. These architectures use *Self-Attention Mechanisms* (SAM) to explore data organised in simple structures, such as sequence relations in texts and neighbourhood relations in image segments, to build general, goal-oriented relations. However, sophisticated data organisation such as graphs [11], has proved challenging for Transformers. Graph Transformers (GT) were developed to address this based on purpose-oriented forms of graph encoding, resulting in molecular attribute prediction that uses topological and geometric information to represent molecular structure. This is the *Similar Property Principle* [12], an assumption that molecular properties are mostly determined by their 3D structure [13].

GT examples include GROVER and MPG [14], which use *Graph Neural Networks* (GNN) to extract local structural information from molecular graphs and feed the resulting embedding into Transformer layers. Recent approaches have directly integrated graph structural information into Transformers via improved positional encoding [15] and attention maps derived from graph topology [16], as well as using inter-atomic distance as a form of geometric information to be explored in attention maps [17]. Concurrently, various evolutionary Computational Chemistry (*CompChem*) methods have also yielded competitive results for *de novo* molecular generation [18]–[21]. In these methods, the chemical search space is defined via specific molecular encoding, and selection and variation operators defined based on molecular fragments [22]. *CompChem* methods for novel molecular synthesis [23], [24] have used Evolutionary Algorithms (EA) in combination with DL [25] for attribute prediction, increasing the likelihood of generating synthetically viable compounds by using EA based stochastic search accelerated by specific domain knowledge inferred using DL.

To address ongoing research efforts that combine DL and *CompChem* for automated design of environmental sustainable chemical products, this study presents an evolutionary molecular design method combining DL and Multi-Objective Optimisation (MOO) [26]. Specifically, we use Geometric Deep Learning [11] to estimate molecular attribute values in combination with evolutionary MOO based on Information-geometric Optimisation [27]. MOO directs molecular search using multiple objectives to minimise toxicity and maximise synthetic accessibility, which are estimated using DL.

II. Methods

A. Prediction of Molecular Attributes

Prediction of molecular attributes can be summarised as: (1) *Given* a set $\mathcal{A}$ of molecular attributes which can be ascribed to compounds; a *problem space* comprised by a cloud of compounds for which 3D structural information and chemical composition are assumed to be available; and one

particular attribute $A \in \mathcal{A}$ whose value is known for a subset of compounds in the problem space; (2) *Find* an appropriate disposition of compounds in an appropriate N-dimensional space such that the distance between pairs of compounds characterises their similarity; and predictions of values of the attribute A for all compounds in the problem space.

We apply the *Uni-Mol* 3D *GNN* [17] as our molecular attribute prediction method. *Uni-Mol* has demonstrated superior molecular attribute task performance prediction on benchmark data sets [17], using a Transformer-based architecture and *Self-Attention Mechanisms* (SAM) to explore structured data and build goal-oriented relations. Molecules are represented as 3D nodes with atom type and 3D coordinates, with invariant spatial positional encoding and pair-level representation to effectively capture 3D information.

Our adopted molecular attribute prediction method adapts relative positional encoding by utilising Euclidean distances of all atom pairs, followed by a pair-type aware Gaussian kernel [28]. Formally, the *D*-channel positional encoding of atom pair ij is denoted as equations (1), (2) and (3), in which equation (3) is the Gaussian density function:

$$p_{ij} = \{G(A(d_{ij}, t_{ij}; \mathbf{a}, \mathbf{b}), \mu^k, \sigma^k) | k \in [1, D]\} \tag{1}$$

$$A(d, r; \mathbf{a}, \mathbf{b}) = a_r d + b_r \tag{2}$$

$$G(d, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(d-\mu)^2}{2\sigma^2}} \tag{3}$$

In these equations μ, 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. Pair-type is determined by atom pair types ij. $A(d_{ij}, t_{ij}; \mathbf{a}, \mathbf{b})$ is the affine transformation with parameters $\mathbf{a}$, $\mathbf{b}$, and d_{ij} corresponding to its pair-type t_{ij}. To initialise pair-level representation we use spatial positional encoding with atom-to-pair communication using multi-head SAM query-key products. Equation 4 is the update of ij pair representation:

$$q_{ij}^0 = p_{ij} M, q_{ij}^{l+1} = q_{ij}^l + \{\frac{Q_i^{l,h}(K_j^{l,h})^T}{\sqrt{d}} | h \in [1, H]\} \tag{4}$$

In this equation q_{ij}^l is the pair representation of atoms ij in l-th layer, H the number of attention heads, d the dimension of hidden representations and $Q_i^{l,h}(K_j^{l,h})$ 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, making the representation the same shape as multi-head SAM query-key product results. Equation (5) denotes SAM with pair-to-atom communication.

$$Attention(Q_i^{l,h}, K_j^{l,h}, V_j^{l,h}) = softmax(\frac{Q_i^{l,h}(K_j^{l,h})^T}{\sqrt{d}} + q_{ij}^{l-1,h}) V_j^{l,h} \tag{5}$$

In this equation $V_j^{l,h}$ is the j-th atom in the l-th layer h-th head. Uni-Mol lacks the capability to directly output 3D coordinates, which is crucial for tasks that require 3D spatial information. Thus, a SE(3)-Equivariance head is used, enabling direct output of 3D coordinates (equations 6 and 7).

$$\hat{x}_i = x_i + \sum_{j=1}^{n} \frac{(x_i - x_j) c_{ij}}{n}, \tag{6}$$

$$c_{ij} = ReLU((q_{ij}^L - q_{ij}^0)U)W \tag{7}$$

In these equations *n* is the number of atoms, *L* the number of layers in model, $x_i \in R^3$ the input coordinate of *i*-th atom, and $\hat{x}_i \in R^3$ 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.

1) Pre-training: Our pre-training data-set consisted of approximately 19 million molecules, sourced from multiple public data sets. To obtain the 3D conformations, a combination of ETKGD [29] and Merck Molecular Force Field optimisation [30] from RDKit tool [31] was used to randomly generate ten conformations for each molecule. For each molecule, a special atom [CLS] is added to represent the entire molecule, with its coordinate being the centre of all atoms. Two additional heads were used 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 (coordinate prediction), utilises the SE(3)-Equivariance coordinate head to predict the correct coordinates for the atoms with corrupted coordinates.

2) Fine-tuning: 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. Where 3D conformations could not be generated, the molecular graph was used as a 2D conformation. Also, the [CLS] mean representation of all atoms (entire molecule), was used in conjunction with a linear head to fine-tune on downstream tasks.

B. Information-geometric Attribute Optimisation

Broadly, we characterise our MOO problem as: (1) *Given* a set $\mathcal{A}$ of relevant attributes which can be ascribed to specified compounds (assumed to have domains ranging through real-valued intervals); a subset $\mathcal{A}_{opt} \subseteq \mathcal{A}$ of those attributes which must be *optimised* (values must be minimised or maximised); a subset $\mathcal{A}_{constr} \subseteq \mathcal{A}$ of those attributes which define *constraints*, such that for each attribute $A_c \in \mathcal{A}_{constr}$ we have defined two values v_c^{min}, v_c^{max}, $v_c^{min} \leq v_c^{max}$; and a *problem space* comprising compounds considered as candidate solutions for the problem, where all compounds

are published in large, public data-sets such as *PubChem* [32]; (2) *Find* a set of compounds which are *sufficiently good approximations* of compounds that optimise attribute values in $\mathcal{A}_{opt}$ and whose attribute values for those attributes $A_c \in \mathcal{A}_{constr}$ belong to the interval $[v_c^{min}, v_c^{max}]$.

To suitably search the problem space we have used the *Similar Property Principle*. Specifically, we use the *Tanimoto similarity* [33], based on the concept of *molecular fingerprints* [34] based on *features vectors*, listing selected substructures and connections between substructures such that a specific compound can be characterised as a binary vector indicating presence or absence of each feature in the compound. The Tanimoto similarity index between compounds M_a and M_b, is defined by equation 8:

$$T_{a,b} = \frac{c}{a+b+c} \tag{8}$$

In this equation a is the number of features in fingerprint M_a, b the number of features present in fingerprint M_b, and c the number of features present in both fingerprints. Herein, we refer to $NN^A_{a,b}$ as the similarity between compounds M_a and M_b based on the measure developed for attribute A. To search for near-optimal compounds given specified attributes $\mathcal{A}_{opt}$ and $\mathcal{A}_{constr}$, we use an algorithm inspired by *Multi-objective Covariance Matrix Adaptation Evolution Strategy* (MO-CMA-ES) [35], adapted to our non-parametric problem space.

Experiments start with an arbitrary *seed compound* M_0 in the problem space. Given a *generic threshold* $\hat{T}$, we retrieve from the problem space the set of compounds $\mathcal{M}_0 = \{M : T_{f(M_0),f(M)} \geq \hat{T}\}$, where, $f(M_i)$ is the fingerprint of compound M_i, and $M_0 \in \mathcal{M}_0$. For each $M_i \in \mathcal{M}_0$, we check if the constraints defined for attributes in $\mathcal{A}_{constr}$ are satisfied, and build $\tilde{\mathcal{M}}_0 \subseteq \mathcal{M}_0$ containing only the compounds that satisfy all constraints. From these, we build the *Pareto front* of candidate solutions which comprise a Pareto equilibrium considering all attributes in $\mathcal{A}_{opt}$, whose values are estimated based on specialised similarities $NN^{A_p}_{f(M_i)}, M_i \in \mathcal{M}_0, A_p \in \mathcal{A}_{opt}$. Thus, this builds the initial solution set $\mathcal{S}_0 = \{M_i : M_i \in \textit{Pareto front}\}$. Given a specified *population size* λ, we select at random M_{01}, ..., $M_{0\lambda}$ from $\mathcal{S}_0$ and, for each compound, repeat the procedure above to build solution sets $\mathcal{S}_{01}$, ..., $\mathcal{S}_{0\lambda}$, and then the overall solution set $\mathcal{S}_1 = \cup_{j=1}^{\lambda}\{\mathcal{S}_{0j}\}$. This procedure is repeated to build $\mathcal{S}_2, \mathcal{S}_3, \ldots$, until a stability criteria is reached ($\frac{|\mathcal{S}_{k+1}|}{|\mathcal{S}_k|} \approx 1$). To avoid local optima, MO-CMA-ES also includes a *growth factor* $\hat{G} > 1$ for λ (Equations 9, 10):

$$If \frac{|\mathcal{S}_{k+1}|}{|\mathcal{S}_k|} < 1, \lambda \rightarrow \lambda \times \hat{G} \tag{9}$$

$$If \frac{|\mathcal{S}_{k+1}|}{|\mathcal{S}_k|} > 1, \lambda \rightarrow \frac{\lambda}{\hat{G}}. \tag{10}$$

Epoch	Non-Toxic	Toxic	Overall	50/50
1	55.2	86.3	55.6	70.8
2	**73.2**	**72.0**	**73.2**	**72.6**
3	84.2	58.5	83.9	71.4
4	91.3	38.4	90.6	64.9
5	90.3	42.1	89.7	66.2
6	94.6	28.7	93.7	61.6
7	96.0	25.9	95.1	61.0
8	93.8	32.3	93.0	63.1
9	96.1	22.9	95.1	59.5
10	96.2	23.2	95.2	59.7

TABLE I: Method (section II) classification accuracy (%). *Overall:* entire data-set. $50/50$*:* considers combined means for *Non-Toxic* and *Toxic* estimates. Bold highlights highest accuracy for all criteria.

III. Experiments and Results

Our goal is chemical product design for environmental sustainability, thus our optimisation attributes are *toxicity*, to be minimised thus reducing *de novo* product environmental impact, and *synthetic accessibility*, to be maximised to reduce production costs. We predicted the aquatic toxicity of molecules as classified by the *Globally Harmonised System of Classification and Labelling of Chemicals* (*GHS*) [36]. We focused on aquatic toxicity given the many molecules classified according to this attribute. To focus experiments, we limited our analysis to the development of detergents for domestic use, where minimising aquatic toxicity is critical. Aquatic toxicity is subdivided into two attributes: aquatic *acute* toxicity and aquatic *chronic* toxicity. Both indicate potentially lethal effects on the aquatic biome. Aquatic acute toxicity is defined as lethal to aquatic life within 96 hours of constant exposure, while aquatic chronic toxicity is defined as lethal to aquatic life within 28 days.

This study's molecule data-set was provided by a private company[1], and contains 251k molecules. Only two percent of the data-set was previously classified with respect to aquatic toxicity. Hence, our experiments merged acute and chronic toxicity to increase training accuracy to obtain a predictor for unclassified molecules. To ensure robust experimental evaluation, the training data-set was partitioned into training, validation, and testing subsets using a scaffold-split strategy, with an $80-10-10$ ratio, respectively. The scaffold-split approach is based on the molecular scaffold of the compounds and is considered more challenging than a random split strategy. To address the imbalance in the number of toxic and non-toxic molecules, we applied random oversampling to the toxic molecules in training set, resulting in improved method accuracy. We trained 20 replications of the method (with different hyper-parameters), where table I presents test results from the best performing version (highest combination of toxic and non-toxic accuracy, bold in table I).

[1]https://www.smarterx.com/

A. Information-geometric Optimisation

Empirical results, for optimising compound attribute values, are reported considering the following criteria:

(1) Robustness with respect to choice of seed compound M_0: it is expected that similar seed compounds lead to similar optimised compounds. We have selected two groups of compounds, each containing similar compounds where compounds in different groups were dissimilar (given Tanimoto similarity). As expected, compounds within the same group led to the same optimised solution set of compounds, and each group led to a different solution set.

(2) Sensitivity with respect to the search space threshold $\hat{T}$. Results indicate solutions originating from higher $\hat{T}$ values are embedded in those originating from lower $\hat{T}$ values (where lower $\hat{T}$ values induce larger search spaces).

(3) Sensitivity with respect to population size λ: smaller values of λ induce more focused solutions requiring more cycles to reach stability, building solutions further from the seed compound than those gotten with larger λ values. Results indicate that solutions originating from smaller λ values are embedded into those originating from larger λ values.

(4) Sensitivity with respect to attributes to be optimised: more attributes increase solution set sizes, given that more attributes means more difficulty finding dominated solutions. Results indicate that if the sets of optimising attributes A and B are such that $A \subseteq B$, then the corresponding solution sets $\mathcal{S}_A$ and $\mathcal{S}_B$ are such that $\mathcal{S}_A \subseteq \mathcal{S}_B$. Also, the selection of attributes constitutes a useful tool to control the traversal of the search space to find effective solution sets.

Experiments apply our method (section II) to estimate toxicity of compounds in solution set $\mathcal{S}_1$. Specifically, to simultaneously optimise accuracy to identify toxic and non-toxic compounds. Results indicate the only compound to be ruled out of $\mathcal{S}_1$ is `CCCCCCCC(CCCCCCC)COS(=O)(=O)O`, highlighted as underlined text in table 2. We observe that this compound is *on the fringe* of the solution set $\mathcal{S}_1$, featuring the minimal value of *XLogP*. This compound is not registered as toxic in *PubChem*, indicating in high probability it was not laboratory tested, and thus the efficacy of our method for detecting potentially toxic compounds.

Experiments use the *PubChem* [32] database, specifically considering laundry detergents and two defining attributes. First, *XLogP* which measures the ratio between *lipophilicity* and *hydrophilicity* of a compound and when maximised provides us with an indication of low toxicity and good cleaning properties. Second, *Molecular Complexity* which measures the size and structural complexity of a compound and when minimised provides us with an approximate indication of synthetic accessibility (level of difficulty to chemically synthesise a compound where high structural complexity indicates a likely low synthetic accessibility). Additionally, to further gauge the optimisation efficacy of our method, we include a *Molecular Weight* attribute to be minimised (to increase likelihood of high synthetic accessibility). Experiments using our optimisation constraints (section III-A) are described in the following.

1) Sensitivity with respect to M_0: As source compounds (M_0), we used groups described by a patented detergent [37]:

- *Group 1*: Methylhexadecyl hydrogen sulphate isomers:
  `CCCCCCC(C)CCCCCCCCOS(=O)(=O)O`;
  `CCCCCCCCC(C)CCCCCCOS(=O)(=O)O`;
  `CCCCCCCCC(CCCCCC)CCOS(=O)(=O)O`.
- *Group 2*: Methylhexadecanol isomers:
  `CCCCCCC(C)CCCCCCCCO`;
  `CCCCCCCCC(C)CCCCCCO`;
  `CCCCCCCCC(CCCCCC)CCO`.

Stability was reached in solution set $\mathcal{S}_1$ for both groups (tables 2 and 3). These results are also presented as *Complexity versus XLogP* graphs (figure 2, top left, top right) depicting the Pareto fronts of each solution set. These results were obtained with $\hat{T} = 0.98$ and $\lambda = 10$. Solution sets were identical for compounds within each group.

2) Sensitivity with respect to $\hat{T}$: The same experiment (1) was repeated with $\hat{T} = 0.95$ ($\lambda = 10$) for the first compound in *Group 1*. Figure 2 (bottom left) presents results including compounds obtained with $\hat{T} = 0.98$. Results indicate that stability was reached only in solution set $\mathcal{S}_4$. With the exception of a single compound obtained with $\hat{T} = 0.98$, all compounds obtained with $\hat{T} = 0.98$ were also obtained with $\hat{T} = 0.95$. Compounds in the final solution set obtained with $\hat{T} = 0.95$ were significantly different from those obtained with $\hat{T} = 0.98$. The solution set obtained with $\hat{T} = 0.98$ contained 10 compounds, whereas the solution set obtained with $\hat{T} = 0.95$ contained 43 compounds.

3) Sensitivity with respect to λ: Experiment (1) was again repeated with $\lambda = 7$ ($\hat{T} = 0.95$) for the first compound in *Group 1*. Results (figure 2, bottom right), indicate stability was reached only in solution set $\mathcal{S}_6$. With the exception of five compounds obtained with $\lambda = 7$ (not present in solutions gotten with $\lambda = 10$), both solution sets coincided.

4) Sensitivity with respect to attributes: This experiment used the seed compound M_0 as the first compound in *Group 1*, $\hat{T} = 0.98$ and $\lambda = 10$, and the following attributes for optimisation: *XLogP* (to be maximised); *Molecular Complexity* (to be minimised); and *Molecular Weight* (to be minimised). Figure 2 (top left) presents results where seven additional compounds were selected given the additional attribute to be optimised. Figure 1 presents the Pareto front for this case, where additional compounds are highlighted in a different colour.

Compound	XLogP	CP
CCCCCCCC(CCCCCCC)COS(=O)(=O)O	6.6	295
CCCCCCCCCCCCC(C)CCOS(=O)(=O)O	6.8	309
CCCCCCCCC(CCCCCCCC)COS(=O)(=O)O	7.7	319
CCCCCCCCCC(CCCCCCCCC)COS(=O)(=O)O	8.8	344
CCCCCCCCCCC(CCCCCCCCCC) COS(=O)(=O)O	9.9	370
CCCCCCCCCCCC(CCCCCCCCCCC) COS(=O)(=O)O	11.0	395
CCCCCCCCCCCCC(CCCCCCCCCCCC) COS(=O)(=O)O	12.0	421
CCCCCCCCCCCCCC(CCCCCCCCCCCCC) COS(=O)(=O)O	13.1	447
CCCCCCCCCCCCCCC(CCCCCCCCCCCCCC) COS(=O)(=O)O	14.2	473
CCCCCCCCCCCCCCCC(CCCCCCCCCCCCCC) COS(=O)(=O)O	14.7	515

TABLE II: Solution set for all *Group 1* molecules, stabilised after one iteration ($\mathcal{S}_1$). CP: Complexity, *XlogP:* Section III-A.

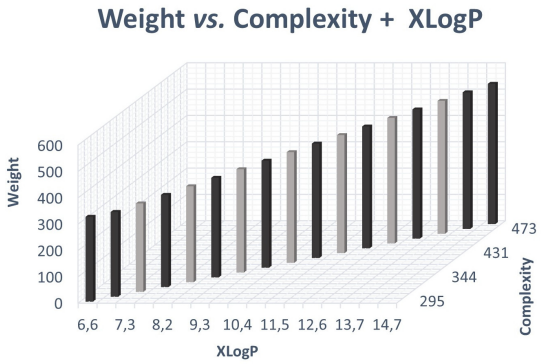

Fig. 1: Pareto front: Compounds highlighted in grey are additional compounds given *Molecular Weight* as another attribute.

Compound	XLogP	CP
CCCCCCCC(CCCCCCC)CO	7.0	120
CCCCCCCC(CCCCCCC)CCO	7.7	131
CCCCCCC(CCCC)CCCCCO	7.8	134
CCCCCCCCC(CCCCCCCC)CO	8.1	140
CCCCCC(CCCC)CCCCCCCO	8.3	145
CCCCCC(CCCCC)CCCCCCCO	8.6	146
CCCCCCCCC(CCCCCCCC)CCO	8.8	151
CCCCCCCCCC(CCCCCCCCC)CO	9.2	161
CCCCCCCCCC(CCCCCCCCC)CCO	9.9	172
CCCCCCCCCCC(CCCCCCCCCC)CO	10.3	182
CCCCCCCCCCC(CCCCCCCCCC)CCO	11.0	194
CCCCCCCCCCCC(CCCCCCCCCCC)CO	11.3	204
CCCCCCCCCCCCCCCCCCCC(C)CCO	11.5	226
CCCCCCCCCCCCC(CCCCCCCCCCC)CO	12.4	227
CCCCCCCCCCCCCC(CCCCCCCCCCCC)CO	13.5	250
CCCCCCCCCCCCCC(CCCCCCCCCCCC)CCO	14.2	262
CCCCCCCCCCCCCCC(CCCCCCCCCCCCC)CO	14.6	273
CCCCCCCCCCCCCCC(CCCCCCCCCCCCC)CCO	15.3	286

TABLE III: Solution set for *Group 2* molecules and stabilised after one iteration ($\mathcal{S}_1$). CP: Complexity (Section III-A.).

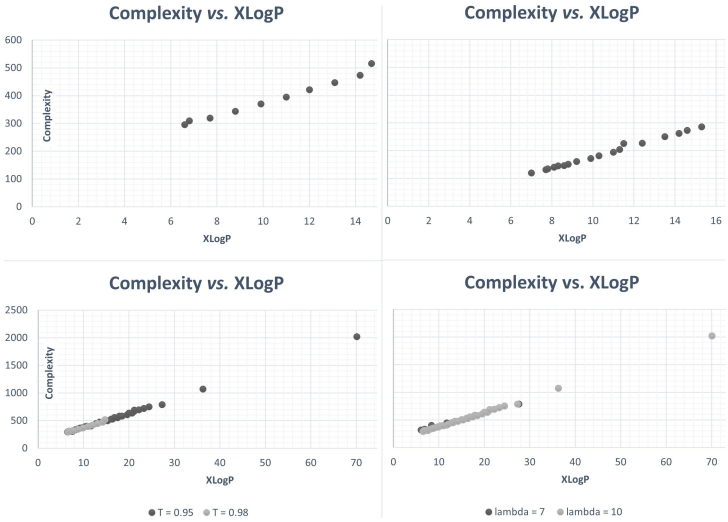

Fig. 2: Top Left: Pareto front for *Group 1* molecules. Top Right: Pareto front for *Group 2* molecules. Lower Left: Pareto front for one molecule in *Group 1*, $\lambda = 10$, $\hat{T} \in \{0.95, 0.98\}$. Lower Right: Pareto front for one molecule in *Group 1*, $\hat{T} = 0.95$, $\lambda \in \{7, 10\}$.

Such results provide expert chemists with the possibility to semi-automate the search process via directing search towards desirable (optimal) compounds. Overall, results indicate that our proposed methods, specifically combining geometric deep-learning and multi-objective optimisation (sections II-A, III-A) for molecular property prediction and optimisation, is a promising approach to provide chemical product designers with a useful computational molecular and optimisation design tool. Results also indicate that the choice of the appropriate, attribute-dependent similarity measures between compounds, seed compound, quantity and quality of attributes to be optimised, and choice of values of hyper-parameters are crucial for the quality of identified *de novo* compounds (section III-A). Specifically, we have identified preliminary empirical evidence that frequently used molecular descriptions such as SMILES can be too coarse to enable proper classification of molecules according to attributes of interest. We applied *Uni-Mol* [17] to estimate aquatic toxicity of compounds, obtaining accuracy considered acceptable by domain experts. *Uni-Mol* yielded an accuracy (table I) comparable with the accuracy and reliability of existing methods [2], [38]–[40], where such methods remain disadvantaged by costly trial and error optimisation.

Computational results analysis also identified at least one compound considered potentially toxic to the environment (section III-A1). However, this study's main contribution was combining *Uni-Mol* with MO-CMA-ES to demonstrate a novel computational tool for searching a chemical space for optimal compounds (in a search space characterised by compounds with partially known attribute values).

IV. Conclusions and Future Work

This study introduced a new computational chemistry tool to automate molecular design while satisfying multiple constraints of new molecular compounds, where molecular attribute values are only partially known. Specifically, our method combined *Uni-Mol*, a GNN to estimate attribute values of compounds given their 3D molecular structure, and MO-CMA-ES, a multi-objective evolutionary search method that delivers a Pareto optimal solution set of compounds. Results indicated the efficacy of this method for discovering new compounds with optimised attributes (toxicity and synthetic accessibility in this study). As future work, we plan further empirical validation of this method as a research support tool for *de novo* chemical synthesis, and potential application to automated synthesis of novel molecular compounds.

References

[1] S. Kearnes, "Pursuing a Prospective Perspective," *TRECHEM*, vol. 3, no. 2, pp. 77–79, 2021.

[2] G. Schneider, "Automating Drug Discovery," *Nature Reviews Drug Discovery*, vol. 17, no. 2, pp. 97–113, 2018.

[3] K. Ng and R. Gani, "Chemical Product Design: Advances in and Proposed Directions for Research and Teaching," *Computers & Chemical Engineering*, vol. 126, pp. 147–156, 2019.

[4] B. Goldman and et al., "Defining Levels of Automated Chemical Design," *Journal of Medicinal Chemistry*, vol. 65, pp. 7073–7087, 2020.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 1, pp. 436–444, 2015.

[6] E. Gawehn, J. Hiss, and G. Schneider, "Deep Learning in Drug Discovery," *Molecular Informatics*, vol. 35, no. 1, pp. 3–14, 2016.

[7] C. Bilodeau, W. Jin, T. Jaakkola, R. Barzilay, and K. F. Jensen, "Generative models for molecular discovery: Recent advances and challenges," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 12, no. 5, p. e1608, 2022.

[8] D. Weininger, "SMILES: A Chemical Language and Information-System," *Journal of Chemical Information & Modeling*, vol. 28, pp. 31–36, 1988.

[9] R. Gomez-Bombarelli and et al., "Automatic Chemical Design using a Data-driven Continuous Representation of Molecules," *ACS Central Science*, vol. 4, no. 2, pp. 268–276, 2018.

[10] A. Zhang, Z. Lipton, M. Li, and A. Smola, "Dive into Deep Learning," *arXiv preprint:2106.11342*, 2021.

[11] M. Bronstein, J. Bruna, T. Cohen, and P. Velickovic, "Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges," *arXiv preprint:2104.13478*, 2021.

[12] J. Mitchell, "Machine Learning Methods in Chemoinformatics," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 4, no. 5, pp. 468–481, 2014.

[13] C. Hansch and T. Fujita, "p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure," *Journal of the American Chemical Society*, vol. 86, no. 8, pp. 1616–1626, 1964.

[14] P. Li and et al., "An effective self-supervised framework for learning expressive molecular global representations to drug discovery," *Briefings in Bioinformatics*, vol. 22, no. 6, 2021.

[15] D. Cai and W. Lam, "Graph Transformer for Graph-to-Sequence Learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, (New York, USA), pp. 7464–7471, AAAI, 2020.

[16] C. Ying and et al., "Do Transformers Really Perform Badly for Graph Representation?," *Advances in Neural Information Processing Systems*, vol. 34, pp. 28877–28888, 2021.

[17] G. Zhou and et al., "Uni-Mol: A Universal 3D Molecular Representation Learning Framework," in *Proceedings of the Eleventh International Conference on Learning Representations*, (Kigali, Rwanda), ICLR, 2023.

[18] N. Yoshikawa and et al., "Population-based De Novo Molecule Generation using Grammatical Evolution," *Chemistry Letters*, vol. 47, no. 11, pp. 1431–1434, 2018.

[19] J. Jensen, "A Graph-based Genetic Algorithm and Generative Monte Carlo Tree Search for the Exploration of Chemical Space," *Chemical Science*, vol. 10, no. 12, pp. 3567–3572, 2019.

[20] Y. Kwon and et al., "Evolutionary Design of Molecules based on Deep Learning and a Genetic Algorithm," *Nature Scientific Reports*, vol. 11, no. 17304, pp. 4–6, 2021.

[21] D. Varela and J. Santos, "Niching Methods Integrated with a Differential Evolution Memetic Algorithm for Protein Structure Prediction," *Swarm and Evolutionary Computation*, vol. 71, no. 1, p. 101062, 2022.

[22] P. Polishchuk, "CReM: Chemically Reasonable Mutations Framework for Structure Generation," *Journal of Cheminformatics*, vol. 12, no. 1, p. 28, 2020.

[23] J. Behler, "Neural Network Potential-energy Surfaces in Chemistry: A Tool for Large-scale Simulations," *Physical Chemistry Chemical Physics*, vol. 13, no. 1, p. 17930, 2011.

[24] M. Reveil and P. Clancy, "Classification of Spatially Resolved Molecular Fingerprints for Machine Learning Applications and Development of a Codebase for their Implementation," *Molecular Systems Design & Engineering*, vol. 3, no. 1, pp. 431–441, 2018.

[25] T. Le and D. Winkler, "Discovery and Optimization of Materials using Evolutionary Approaches," *Chemical Reviews*, vol. 116, no. 1, pp. 6107–6132, 2016.

[26] R. Winter and et al., "Efficient Multi-objective Molecular Optimization in a Continuous Latent Space," *Chemical Science*, vol. 10, no. 34, pp. 8016–8024, 2019.

[27] Y. Ollivier and et al., "Information-geometric Optimization Algorithms: A Unifying Picture via Invariance Principles," *Journal of Machine Learning Research*, vol. 18, no. 18, pp. 1–65, 2017.

[28] M. Shuaibi and et al., "Rotation Invariant Graph Neural Networks using Spin Convolutions," *arXiv preprint:2106.09575*, 2021.

[29] S. Riniker and G. Landrum, "Better Informed Distance Geometry: Using what we Know to Improve Conformation Generation," *Journal of Chemical Information & Modeling*, vol. 55, no. 12, pp. 2562–2574, 2015.

[30] T. Halgren, "Merck Molecular Force Field," *Journal of computational chemistry*, vol. 17, no. 5-6, pp. 490–519, 1996.

[31] G. Landrum and et al., "RDKit: A Software Suite for Cheminformatics, Computational Chemistry and Predictive Modeling," *Greg Landrum*, vol. 8, 2013.

[32] S. Kim and et al., "PubChem Substance and Compound Databases," *Nucleic Acids Research*, vol. 44, pp. 1202–1213, 2016.

[33] D. Bajusz, A. Rácz, and K. Héberger, "Why is Tanimoto Index an Appropriate Choice for Fingerprint-based Similarity Calculations?," *Journal of Cheminformatics*, vol. 7, no. 1, pp. 1–13, 2015.

[34] A. Cereto-Massagué and et al., "Molecular Fingerprint Similarity Search in Virtual Screening," *Methods*, vol. 71, pp. 58–63, 2015.

[35] C. Igel, N. Hansen, and S. Roth, "Covariance Matrix Adaptation for Multi-objective Optimization," *Evolutionary computation*, vol. 15, no. 1, pp. 1–28, 2007.

[36] C. Winder, R. Azzi, and D. Wagner, "The Development of the Globally Harmonized System (GHS) of Classification and Labelling of Hazardous Chemicals," *Journal of Hazardous Materials*, vol. 125, no. 1-3, pp. 29–44, 2005.

[37] R. Ellison and et al., "Laundry Detergent Composition and Method of Making Thereof," June 18 2015. US Patent App. 14/402,327.

[38] G. Schneider and U. Fechner, "Computer-based de novo Design of Drug-like Molecules," *Nature Reviews Drug Discovery*, vol. 4, no. 1, pp. 649–663, 2005.

[39] R. Chen and et al., "Machine Learning for Drug-Target Interaction Prediction," *Molecules*, vol. 23, no. 9, p. 2208, 2018.

[40] L. Pu and et al., "eToxPred: A Machine Learning-based Approach to Estimate the Toxicity of Drug Candidates," *BMC Pharmacology and Toxicology*, vol. 20, no. 2, pp. 1–15, 2019.