International Journal of
Molecular Sciences

MDPI

Article

An Inverse QSAR Method Based on a Two-Layered Model and Integer Programming

Yu Shi [1], Jianshen Zhu [1,†], Naveed Ahmed Azam [1,†], Kazuya Haraguchi [1,†], Liang Zhao [2], Hiroshi Nagamochi [1,*,†] and Tatsuya Akutsu [3]

[1] Department of Applied Mathematics and Physics, Kyoto University, Kyoto 606-8501, Japan; shi@amp.i.kyoto-u.ac.jp (Y.S.); zhujs@amp.i.kyoto-u.ac.jp (J.Z.); azam@amp.i.kyoto-u.ac.jp (N.A.A.); haraguchi@amp.i.kyoto-u.ac.jp (K.H.)

[2] Graduate School of Advanced Integrated Studies in Human Survivability (Shishu-Kan), Kyoto University, Kyoto 606-8306, Japan; liang@gsais.kyoto-u.ac.jp

[3] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji 611-0011, Japan; takutsu@kuicr.kyoto-u.ac.jp

* Correspondence: nag@amp.i.kyoto-u.ac.jp; Tel.: +81-753-4920

† These authors contributed equally to this work.

Abstract: A novel framework for inverse quantitative structure–activity relationships (inverse QSAR) has recently been proposed and developed using both artificial neural networks and mixed integer linear programming. However, classes of chemical graphs treated by the framework are limited. In order to deal with an arbitrary graph in the framework, we introduce a new model, called a two-layered model, and develop a corresponding method. In this model, each chemical graph is regarded as two parts: the exterior and the interior. The exterior consists of maximal acyclic induced subgraphs with bounded height, the interior is the connected subgraph obtained by ignoring the exterior, and the feature vector consists of the frequency of adjacent atom pairs in the interior and the frequency of chemical acyclic graphs in the exterior. Our method is more flexible than the existing method in the sense that any type of graphs can be inferred. We compared the proposed method with an existing method using several data sets obtained from PubChem database. The new method could infer more general chemical graphs with up to 50 non-hydrogen atoms. The proposed inverse QSAR method can be applied to the inference of more general chemical graphs than before.

Keywords: QSAR; molecular design; artificial neural network; mixed integer linear programming; enumeration of graphs; cheminformatics; materials informatics

check for updates

Citation: Shi, Y.; Zhu, J.; Azam, N. A.; Haraguchi, K.; Zhao, L.; Nagamochi, H.; Akutsu, T. An Inverse QSAR Method Based on a Two-Layered Model and Integer Programming. *Int. J. Mol. Sci.* **2021**, *22*, 2847. https://doi.org/10.3390/ijms22062847

Academic Editor: Hanoch Senderowitz

Received: 19 February 2021
Accepted: 7 March 2021
Published: 11 March 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Computer-aided design of chemical structures is one of the key topics in chemoinformatics. In particular, extensive studies have been done on inverse quantitative structure–activity relationships (inverse QSAR), which seek chemical structures having desired chemical activities under some constraints. In this framework, chemical compounds are usually represented as vectors of real or integer numbers, which are often called *descriptors* in chemoinformatics and correspond to *feature vectors* in machine learning. Using these chemical descriptors, various heuristic and statistical methods have been developed for inverse QSAR [1–3]. In many of such methods, inference or enumeration of graph structures from a given set of descriptors is a crucial subtask. Although various methods have been developed for that purpose [4–7], enumeration still remains a challenging task because the number of possible chemical graphs is huge, for example, chemical graphs with up to 30 atoms (vertices) C, N, O, and S, may exceed 10^{60} [8]. Furthermore, even inference is a challenging task because it is NP-hard (computationally difficult) except for some simple cases [9]. Due to this inherent difficulty, most existing methods for inverse QSAR do not guarantee optimal or exact solutions.

On the other hand, the design of novel graph structures has recently become a hot topic in artificial neural network (ANN) studies, and thus extensive studies have been done for inverse QSAR using ANNs, especially with graph convolutional networks [10]. For example, variational autoencoders [11], recurrent neural networks [12,13], grammar variational autoencoders [14], generative adversarial networks [15], and invertible flow models [16,17] have been applied. Note that QSAR using three-dimensional structures of chemical compounds (3D-QSAR) has also been studied [18]. Particularly, comparative molecular field analysis (CoMFA) has been extensively studied and applied to various molecular design problems [19,20]. In CoMFA, electrostatic potential interaction energies across superimposed molecular structures are used as descriptors and then regression is performed by using the partial least squares (PLS) fitting. Recently, deep neural networks have been applied to 3D-QSAR by combining potential interaction energies with convolutional neural networks [21]. However, in order to apply 3D-QSAR, we need to calculate accurate three-dimensional structures of chemical compounds, which is not a straightforward task.

A novel framework for inferring chemical graphs has recently been developed [22,23] based on ANNs and mixed integer linear programming (MILP), as illustrated in Figure 1. It constructs a prediction function in the first phase and infers a chemical graph in the second phase. The first phase of the framework consists of three stages. In Stage 1, we choose a chemical property π and a class $\mathcal{G}$ of graphs, where a property function a is defined so that $a(G)$ is the value of π in $G \in \mathcal{G}$, and collect a data set D_π of chemical graphs in $\mathcal{G}$ such that $a(G)$ is available. In Stage 2, we introduce a feature function $f : \mathcal{G} \to \mathbb{R}^K$ for a positive integer K. In Stage 3, we construct a prediction function $\eta_\mathcal{N}$ with an ANN $\mathcal{N}$ that, given a vector $x \in \mathbb{R}^K$, returns a value $y = \eta_\mathcal{N}(x) \in \mathbb{R}$ so that $\eta_\mathcal{N}(f(G))$ serves as a predicted value to $a(G)$ for each $G \in D_\pi$. Given a target chemical value y^*, the second phase infers chemical graphs G^* with $\eta_\mathcal{N}(f(G^*)) = y^*$ in the next two stages. In Stage 4, we formulate an MILP that simulates the construction of $f(G)$ from G and the computation process in the ANN so that given a target value, y^*, and solve the MILP to infer a chemical graph $G^\dagger$ and a feature vector x^* such that $f(G^\dagger) = x^*$ and $\eta_\mathcal{N}(x^*) = y^*$. In Stage 5, we generate other chemical graphs G^* such that $\eta_\mathcal{N}(f(G^*)) = y^*$ based on the output chemical graph $G^\dagger$.

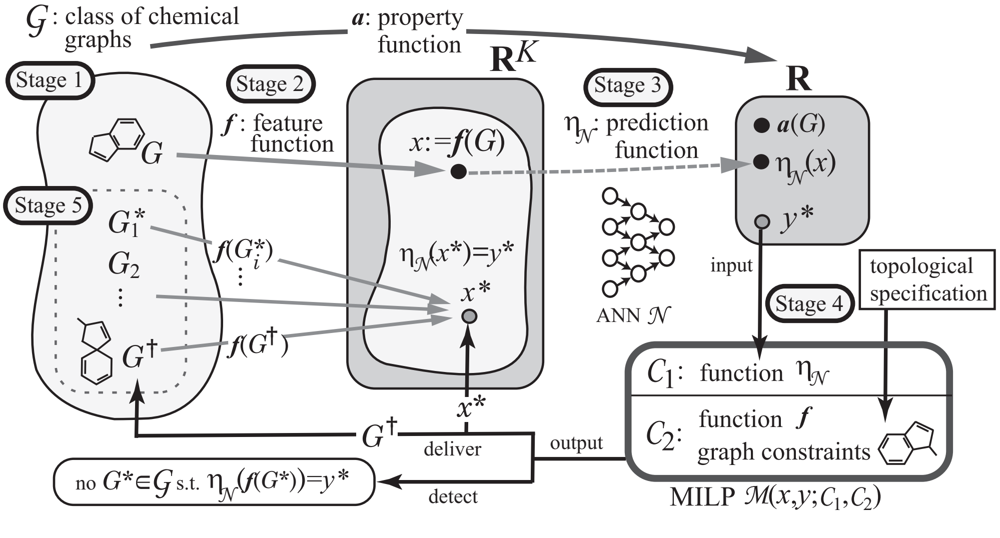


Figure 1. An illustration of a framework for inferring a set of chemical graphs G^*.

MILP formulations required in Stage 4 have been designed for chemical compounds with cycle index 0 (i.e., acyclic) [23,24], cycle index 1 [25], and cycle index 2 [26]. In particular, Azam et al. [24] introduced a restricted class of acyclic graphs that is characterize by an integer ρ, called a "branch-parameter" such that the restricted class still covers most of the acyclic chemical compounds in the database.

Recently, Akutsu and Nagamochi [27] extended the idea to define a restricted class of cyclic graphs, called "ρ-lean cyclic graphs", that covers most of the cyclic chemical compounds in the database. Based on this, they also defined a set of rules for specifying several topological substructures of a target chemical graph in a flexible way in Stage 4 before we solve an MILP. The method has been implemented by Zhu et al. [28], and computational results showed that chemical graphs with around up to 50 non-hydrogen atoms can be inferred. Although the method can infer the class of ρ-lean cyclic graphs and specify topological structures of the cyclic part, we still need to introduce a new model to deal with an arbitrary graph and to include a prescribed structure in the acyclic part of a target chemical graph.

In this paper, we introduce a new model, called a *two-layered model*, for representing the feature of a chemical graph in order to deal with an arbitrary graph in the framework. In the two-layered model, a chemical graph G with a parameter $\rho \geq 1$ is regarded as two parts: the exterior and the interior. The exterior consists of maximal acyclic induced subgraphs with height at most ρ and the interior is the connected subgraph obtained by ignoring the exterior. We define a feature vector $f(G)$ of a chemical graph G to be the frequency of adjacent atom pairs in the interior and the frequency of chemical acyclic graphs in the exterior. Figure 2 illustrates an example of a chemical graph G. For a branch-parameter $\rho = 2$, the interior of the chemical graph G in Figure 2 is obtained by removing the set of vertices with degree 1 $\rho = 2$ times, i.e., first remove the set $V_1 = \{w_1, w_2, \ldots, w_{14}\}$ of vertices of degree 1 in G, and then remove the set $V_2 = \{w_{15}, w_{16}, \ldots, w_{19}\}$ of vertices of degree 1 in $G - V_1$, where the removed vertices become the exterior-vertices of G and there are eight rooted trees $T_1, T_2, \ldots, T_8$ in the exterior of G.

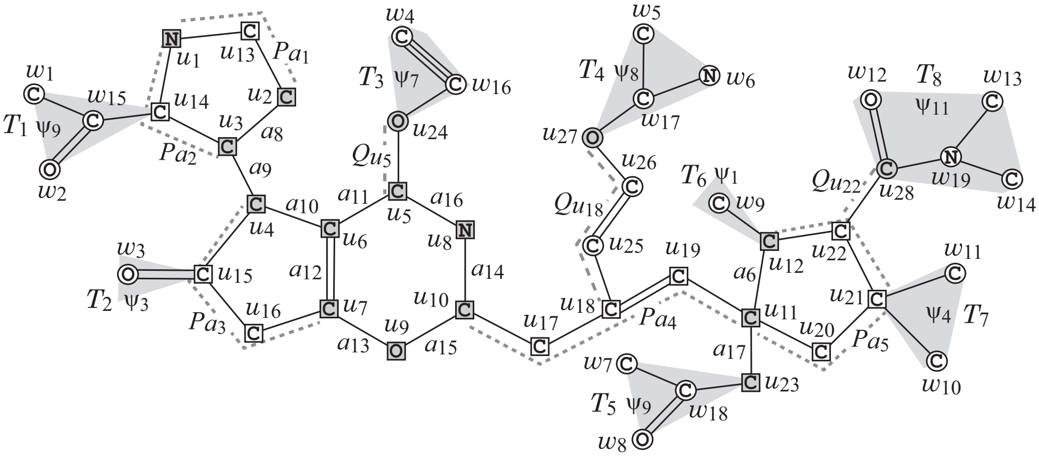
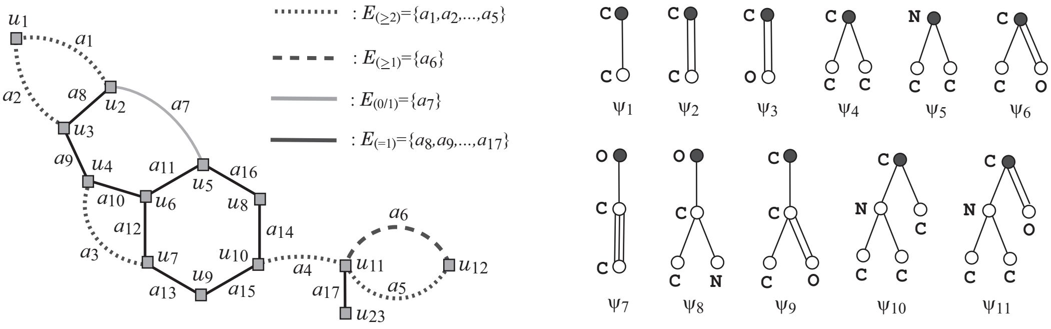


Figure 2. An illustration of a chemical graph G, where for $\rho = 2$, the exterior-vertices are $w_1, w_2, \ldots, w_{19}$ and the interior-vertices are $u_1, u_2, \ldots, u_{28}$.

We also introduce a new set of rules for specifying topological substructures of a target chemical graph G to be inferred so that a prescribed structure can be included in both of the acyclic and cyclic parts of G. The set of rules contains (i) a *seed graph* G_{C} as an abstract form of a target chemical graph G; (ii) a set $\mathcal{F}$ of chemical rooted trees as candidates for trees in the exterior of G; and (iii) lower and upper bounds on the number of components in a target chemical graph such as chemical elements, double/triple bounds and the interior-vertices in G. Figure 3a,b illustrates examples of a seed graph G_{C} and a set $\mathcal{F}$ of chemical rooted trees, respectively. Given a seed graph G_{C}, the interior of a target chemical graph G is constructed from G_{C} by replacing some edges $a = uv$ with paths P_a between the end-vertices u and v, and by attaching new paths Q_v to some vertices v. For example, the chemical graph G in Figure 2 is constructed from the seed graph G_{C} in Figure 3a as follows. First replace five edges $a_1 = u_1u_2, a_2 = u_1u_3, a_3 = u_4u_7, a_4 = u_{10}u_{11}$ and $a_5 = u_{11}u_{12}$ in G_{C} with new paths $P_{a_1} = (u_1, u_{13}, u_2)$, $P_{a_2} = (u_1, u_{14}, u_3)$, $P_{a_3} = (u_4, u_{15}, u_{16}, u_7)$, $P_{a_4} = (u_{10}, u_{17}, u_{18}, u_{19}, u_{11})$ and $P_{a_5} = (u_{11}, u_{20}, u_{21}, u_{22}, u_{12})$, respectively, to obtain the subgraph G_1 of G that consists of vertices depicted with squares. Next, attach to this graph G_1 three new paths, $Q_{u_5} = (u_5, u_{24})$, $Q_{u_{18}} = (u_{18}, u_{25}, u_{26}, u_{27})$, and $Q_{u_{22}} = (u_{22}, u_{28})$, to obtain the interior of G in Figure 2. Finally, the chemical graph G in Figure 2 is obtained by attaching eight trees $T_1, T_2, \dots, T_8$ selected from the set $\mathcal{F}$ and assigning chemical elements and bond-multiplicities in the interior. The frequency of chemical elements and the graph size are controlled with lower and upper bounds on the components in a target chemical graph G. See Section 2.2 for more details on the specification.

We implemented the two-layered model and the results of computational experiments suggest that the proposed method can infer chemical graphs with around up to 50 non-hydrogen atoms.



(a) A seed graph $G_C=(V_C,E_C)$ (b) A set $\mathcal{F}$ of chemical rooted trees

Figure 3. (**a**) An illustration of a seed graph G_C where the vertices in V_C are depicted with gray squares, the edges in $E_{(\geq 2)}$ are depicted with dotted lines, the edges in $E_{(\geq 1)}$ are depicted with dashed lines, the edges in $E_{(0/1)}$ are depicted with gray bold lines, and the edges in $E_{(=1)}$ are depicted with black solid lines. (**b**) A set $\mathcal{F} = \{\psi_1, \psi_2, \dots, \psi_{11}\} \subseteq \mathcal{F}(D_\pi)$ of 11 chemical rooted trees $\psi_i, i \in [1, 11]$, where the root of each tree is depicted with a black circle.

The paper is organized as follows. Section 2.1 introduces some notions on graphs, a modeling of chemical compounds, and a choice of descriptors. Section 2.2 introduces a method of specifying topological substructures of target chemical graphs in Stage 4. Section 3 reports the results on some computational experiments conducted for chemical properties such as octanol/water partition coefficient, boiling point, melting point, flash point, lipophylicity, and solubility. Section 4 makes some concluding remarks. An MILP formulation used in Stage 4 and a review of the dynamic programming algorithm for generating isomers in Stage 5 can be found in Supplementary Materials. The proposed method/system is available at GitHub https://github.com/ku-dml/mol-infer.

2. Materials and Methods

This section presents mathematical details of our developed method. Readers not interested in mathematical details can skip this section.

2.1. Preliminary

This section introduces some notions and terminology on graphs, a modeling of chemical compounds, and our choice of descriptors.

Let $\mathbb{R}$, $\mathbb{Z}$ and $\mathbb{Z}_+$ denote the sets of reals, integers and non-negative integers, respectively. For two integers a and b, let $[a, b]$ denote the set of integers i with $a \leq i \leq b$.

Graphs. Given a graph G, let $V(G)$ and $E(G)$ denote the sets of vertices and edges, respectively. For a subset $V' \subseteq V(G)$ (resp., $E' \subseteq E(G)$) of a graph G, let $G - V'$ (resp., $G - E'$) denote the graph obtained from G by removing the vertices in V' (resp., the edges in E'), where we remove all edges incident to a vertex in V' in $G - V'$. The *rank* $\mathrm{r}(G)$ of a graph G is defined to be the minimum $|F|$ of an edge subset $F \subseteq E(G)$ such that $G - F$ contains no cycle. A path with two end-vertices u and v is called a *u, v-path*. An edge $e = u_1u_2$ in a connected graph G is called a *bridge* if the graph $G - e$ obtained from G by removing edge e is not connected, i.e., $G - e$ consists of two connected graphs G_i containing vertex u_i, $i = 1, 2$. For a cyclic graph G, an edge e is called a *core-edge* if it is in a cycle of G or is a bridge $e = u_1u_2$ such that each of the connected graphs G_i, $i = 1, 2$ of $G - e$ contains a cycle. A vertex incident to a core-edge is called a *core-vertex* of G.

A vertex designated in a graph G is called a *root*. In this paper, we designated at most two vertices as roots, and denote by $\mathrm{Rt}(G)$ the set of roots of G. We call a graph G *rooted* (resp., *bi-rooted*) if $|\mathrm{Rt}(G)| = 1$ (resp., $|\mathrm{Rt}(G)| = 2$), where we call G *unrooted* if $\mathrm{Rt}(G) = \emptyset$.

For a graph G, possibly with roots, a *leaf-vertex* is defined to be a non-root vertex $v \in V(G) \setminus \mathrm{Rt}(G)$ with degree 1, call the edge uv incident to a leaf vertex v a *leaf-edge*, and denote $V_{\mathrm{leaf}}(G)$ and $E_{\mathrm{leaf}}(G)$ the sets of leaf-vertices and leaf-edges in G, respectively. For a graph or a rooted graph G, we define graphs $G_i, i \in \mathbb{Z}_+$ obtained from G by removing the set of leaf-vertices i times so that

$$G_0 := G; \;\; G_{i+1} := G_i - V_{\mathrm{leaf}}(G_i),$$

where we call a vertex $v \in V_{\mathrm{leaf}}(G_k)$ a *leaf k-branch* and we say that a vertex $v \in V_{\mathrm{leaf}}(G_k)$ has height *height* $\mathrm{ht}(v) = k$ in G. The *height* $\mathrm{ht}(T)$ of a rooted tree T is defined to be the maximum of $\mathrm{ht}(v)$ of a vertex $v \in V(T)$. For an integer $k \geq 0$, we call a rooted tree T *k-lean* if T has at most one leaf k-branch. For an unrooted cyclic graph G, we regard the set of non-core-edges in G induces a collection $\mathcal{T}$ of trees each of which is rooted at a core-vertex, where we call G *k-lean* if each of the rooted trees in $\mathcal{T}$ is k-lean. Nearly 97% of cyclic chemical compounds with up to 100 non-hydrogen atoms in PubChem are 2-lean [24].

Two-layered Model. Let G be an unrooted graph. For an integer $\rho \geq 0$, which we call a *branch-parameter*, a *two-layered model* of G is a partition of G into an "interior" and an "exterior" in the following way. We call a vertex $v \in V(G)$ (resp., an edge $e \in E(G)$) of G an *exterior-vertex* (resp., *exterior-edge*) if $\mathrm{ht}(v) < \rho$ (resp., e is incident to an exterior-vertex) and denote the sets of exterior-vertices and exterior-edges by $V^{\mathrm{ex}}(G)$ and $E^{\mathrm{ex}}(G)$, respectively and denote $V^{\mathrm{int}}(G) = V(G) \setminus V^{\mathrm{ex}}(G)$ and $E^{\mathrm{int}}(G) = E(G) \setminus E^{\mathrm{ex}}(G)$, respectively. We call a vertex in $V^{\mathrm{int}}(G)$ (resp., an edge in $E^{\mathrm{int}}(G)$) an *interior-vertex* (resp., *interior-edge*). The set $E^{\mathrm{ex}}(G)$ of exterior-edges forms a collection of connected graphs each of which is regarded as a rooted tree T rooted at the vertex $v \in V(T)$ with the maximum $\mathrm{ht}(v)$, where we call T a *ρ-fringe-tree* (or a fringe-tree). Let $\mathcal{T}^{\mathrm{ex}}(G)$ denote the set of fringe-trees in G. The *interior* of G is defined to be the subgraph $(V^{\mathrm{int}}(G), E^{\mathrm{int}}(G))$ of G. Note that every core-vertex (resp., core-edge) in G is an interior-vertex (resp., interior-edge) of G. Figure 2 illustrates an example of a graph G, such that $V^{\mathrm{int}} = \{u_1, u_2, \ldots, u_{28}\}$, $V^{\mathrm{ex}} = \{w_1, w_2, \ldots, w_{19}\}$ and $\mathcal{T}^{\mathrm{ex}}(G) = \{T_1, T_2, \ldots, T_8\}$ for a branch-parameter $\rho = 2$.

2.1.1. Modeling of Chemical Compounds

To represent a chemical compound, we assume that each chemical element $\mathtt{a}$ has a unique valence $\mathrm{val}(\mathtt{a}) \in [1, 4]$ and we use a hydrogen-suppressed model, because hydrogen atoms can be added at the final stage under the assumption. In the hydrogen-suppressed model, a chemical compound C is represented by a tuple $G = (H, \alpha, \beta)$ of a simple, connected undirected graph H and functions $\alpha : V(H) \to \Lambda$ and $\beta : E(H) \to [1, 3]$, where Λ is a set of non-hydrogen chemical elements such as $\mathtt{C}$ (carbon), $\mathtt{O}$ (oxygen), $\mathtt{N}$ (nitrogen), and so on. The set of atoms and the set of bonds in the compound C are represented by the vertex set $V(H)$ and the edge set $E(H)$, respectively. The chemical element assigned to a vertex $v \in V(H)$ is represented by $\alpha(v)$ and the bond-multiplicity between two adjacent vertices $u, v \in V(H)$ is represented by $\beta(e)$ of the edge $e = uv \in E(H)$. We say that two tuples $(H_i, \alpha_i, \beta_i), i = 1, 2$ are *isomorphic* if they admit an isomorphism ϕ, i.e., a bijection $\phi : V(H_1) \to V(H_2)$ such that $uv \in E(H_1), \alpha_1(u) = \mathtt{a}, \alpha_1(v) = \mathtt{b}, \beta_1(uv) = m \leftrightarrow \phi(u)\phi(v) \in E(H_2), \alpha_2(\phi(u)) = \mathtt{a}, \alpha_2(\phi(v)) = \mathtt{b}, \beta_2(\phi(u)\phi(v)) = m$. When H_i is rooted at a vertex $r_i, i = 1, 2$, $(H_i, \alpha_i, \beta_i), i = 1, 2$ are *rooted-isomorphic* (r-isomorphic) if they admit an isomorphism ϕ such that $\phi(r_1) = r_2$. Chemical rooted trees T_1 and T_5 in Figure 2 are r-isomorphic.

Associated with the two functions α and β in a tuple $G = (H, \alpha, \beta)$, we introduce the following functions: $\beta_G : V(H) \to [0, 12]$, $\mathrm{ac} : V(E) \to \Lambda \times \Lambda \times [1, 3]$, $\mathrm{cs} : V(E) \to \Lambda \times [1, 4]$, and $\mathrm{ec} : V(E) \to (\Lambda \times [1, 4]) \times (\Lambda \times [1, 4]) \times [1, 3]$.

For a notational convenience, we use a function $\beta_G : V(H) \to [0, 4]$ such that $\beta_G(u)$ means the sum of bond-multiplicities of edges incident to a vertex u, i.e.,

$$\beta_G(u) \triangleq \sum_{uv \in E(H)} \beta(uv) \text{ for each vertex } u \in V(H).$$

A *chemical graph* G is defined to be a tuple (H, α, β) such that the valence condition at each vertex $v \in V(H)$ is satisfied, i.e.,

$$\beta_G(v) \le \mathrm{val}(\alpha(v)),$$

where we define the *hydro-degree* $\deg_{\mathrm{hyd}}(v)$ of a vertex v to be $\mathrm{val}(\alpha(v)) - \beta_G(v)$.

Figure 2 illustrates an example of a chemical graph $G = (H, \alpha, \beta)$.

To represent a feature of an edge $e = uv \in E(H)$ such that $\alpha(u) = \mathtt{a}$, $\alpha(v) = \mathtt{b}$ and $\beta(e) = m$ in a chemical graph $G = (H, \alpha, \beta)$, we use a tuple $(\mathtt{a}, \mathtt{b}, m) \in \Lambda \times \Lambda \times [1, 3]$, which we call the *adjacency-configuration* $\mathrm{ac}(e)$ of the edge e. We introduce a total order $<$ over the elements in Λ to distinguish with $(\mathtt{a}, \mathtt{b}, m)$ and $(\mathtt{b}, \mathtt{a}, m)$ $(\mathtt{a} \ne \mathtt{b})$ notationally. For a tuple $\nu = (\mathtt{a}, \mathtt{b}, m)$, let $\overline{\nu}$ denote the tuple $(\mathtt{b}, \mathtt{a}, m)$.

To represent a feature of a vertex $v \in V(H)$ with $\alpha(v) = \mathtt{a}$ that has d atoms in its neighbor in a chemical graph $G = (H, \alpha, \beta)$, we use a pair $(\mathtt{a}, d) \in \Lambda \times [1, 4]$, which we call the *chemical symbol* $\mathrm{cs}(v)$ of the vertex v. We treat $(\mathtt{a}, d)$ as a single symbol $\mathtt{a}d$, and define Λ_{dg} to be the set of all chemical symbols $\mu = \mathtt{a}d \in \Lambda \times [1, 4]$.

To represent a feature of an edge $e = uv \in E(H)$ such that $\mathrm{cs}(u) = \mu$, $\mathrm{cs}(v) = \xi$ and $\beta(e) = m$ in a chemical graph $G = (H, \alpha, \beta)$, we use a tuple $(\mu, \xi, m) \in \Lambda_{\mathrm{dg}} \times \Lambda_{\mathrm{dg}} \times [1, 3]$, which we call the *edge-configuration* $\mathrm{ec}(e)$ of the edge e. We introduce a total order $<$ over the elements in Λ_{dg} to distinguish with (μ, ξ, m) and (ξ, μ, m) $(\mu \ne \xi)$ notationally. For a tuple $\gamma = (\mu, \xi, m)$, let $\overline{\gamma}$ denote the tuple (ξ, μ, m).

To represent a feature of the exterior of a chemical graph $G = (H, \alpha, \beta)$, a ρ-fringe-tree in $\mathcal{T}^{\mathrm{ex}}(G)$ is called a *fringe-configuration* in the exterior.

2.1.2. Introducing Descriptors of Feature Vectors

This section introduces descriptors to define our feature vectors. Let π be a chemical property for which we will construct a prediction function $\eta_{\mathcal{N}}$ from a feature vector $f(G)$ of a chemical graph to a predicted value $y \in \mathbb{R}$ for the chemical property of G.

We first choose a set Λ of non-hydrogen chemical elements and then collect a data set D_π of chemical compounds C whose chemical elements belong to $\Lambda \cup \{\mathtt{H}\}$, where we regard D_π as a set of chemical graphs that represent the chemical compounds C in D_π. To define the interior/exterior of chemical graphs $G \in D_\pi$, we next choose a branch-parameter ρ, where we recommend $\rho = 2$.

Let $\Lambda^{\mathrm{int}}(D_\pi)$ (resp., $\Lambda^{\mathrm{ex}}(D_\pi)$) denote the set of chemical elements used in the set of interior-vertices (resp., exterior-vertices) over all chemical graphs $G \in D_\pi$, and $\Gamma^{\mathrm{int}}(D_\pi)$ denote the set of edge-configurations used in the set of interior-edges over all chemical graphs $G \in D_\pi$. Let $\mathcal{F}(D_\pi)$ denote the set of chemical rooted trees ψ r-isomorphic to a ρ-fringe-tree $T \in \mathcal{T}^{\mathrm{ex}}(G)$ over all chemical graphs $G \in D_\pi$.

We define an integer encoding of a finite set A of elements to be a bijection $\sigma : A \to [1, |A|]$, where we denote by $[A]$ the set $[1, |A|]$ of integers. Introduce an integer coding of each of the sets $\Lambda^{\mathrm{int}}(D_\pi)$, $\Lambda^{\mathrm{ex}}(D_\pi)$, $\Gamma^{\mathrm{int}}(D_\pi)$ and $\mathcal{F}(D_\pi)$. Let $[\mathtt{a}]^{\mathrm{int}}$ (resp., $[\mathtt{a}]^{\mathrm{ex}}$) denote the coded integer of an element $\mathtt{a} \in \Lambda^{\mathrm{int}}(D_\pi)$ (resp., $\mathtt{a} \in \Lambda^{\mathrm{ex}}(D_\pi)$), $[\gamma]$ denote the coded integer of an element γ in $\Gamma^{\mathrm{int}}(D_\pi)$ and $[\psi]$ denote an element ψ in $\mathcal{F}(D_\pi)$.

For each chemical element $\mathtt{a} \in \Lambda$, let $\mathrm{mass}(\mathtt{a})$ and $\mathrm{val}(\mathtt{a})$ denote the mass and valence of $\mathtt{a}$, respectively. In our model, we use integers $\mathrm{mass}^*(\mathtt{a}) = \lfloor 10 \cdot \mathrm{mass}(\mathtt{a}) \rfloor$, $\mathtt{a} \in \Lambda$.

We define the *feature vector* $f(G)$ of a chemical graph $G = (H, \alpha, \beta) \in D_\pi$ to be a vector that consists of the following non-negative integer descriptors $\mathrm{dcp}_i(G)$, $i \in [1, K]$, where $K = 17 + |\Lambda^{\mathrm{int}}(D_\pi)| + |\Lambda^{\mathrm{ex}}(D_\pi)| + |\Gamma^{\mathrm{int}}(D_\pi)| + |\mathcal{F}(D_\pi)|$.

1. $\mathrm{dcp}_1(G)$: the number $n(G) = |V(G)|$ of vertices in G.
2. $\mathrm{dcp}_2(G)$: the number $|V^{\mathrm{int}}(G)|$ of interior-vertices in G.
3. $\mathrm{dcp}_3(G)$: the average $\overline{\mathrm{ms}}(G)$ of mass* over all non-hydrogen atoms in G, i.e., $\overline{\mathrm{ms}}(G) \triangleq \sum_{v \in V(G)} \mathrm{mass}^*(\alpha(v))/n(G)$.
4. $\mathrm{dcp}_i(G)$, $i = 3 + d, d \in [1, 4]$: the number $\mathrm{dg}_d(G)$ of interior-vertices of degree d in G.
5. $\mathrm{dcp}_i(G)$, $i = 7 + d, d \in [1, 4]$: the number $\mathrm{dg}_d^{\mathrm{int}}(G)$ of interior-vertices of interior-degree $\deg_{(V^{\mathrm{int}}, E^{\mathrm{int}})}(v) = d$ in the interior $(V^{\mathrm{int}}, E^{\mathrm{int}})$ of G.
6. $\mathrm{dcp}_i(G)$, $i = 11 + d, d \in [0, 3]$: the number $\mathrm{hydg}_d(G)$ of vertices in G of hydro-degree $\deg_{\mathrm{hyd}}(v) = d$.
7. $\mathrm{dcp}_i(G)$, $i = 15 + m$, $m \in [2, 3]$: the number $\mathrm{bd}_m^{\mathrm{int}}(G)$ of interior-edges with bond multiplicity m in G, i.e., $\mathrm{bd}_m^{\mathrm{int}}(G) \triangleq \{e \in E^{\mathrm{int}} \mid \beta(e) = m\}$.
8. $\mathrm{dcp}_i(G)$, $i = 17 + [\mathtt{a}]^{\mathrm{int}}$, $\mathtt{a} \in \Lambda^{\mathrm{int}}(D_\pi)$: the frequency $\mathrm{na}_{\mathtt{a}}^{\mathrm{int}}(G)$ of chemical element $\mathtt{a}$ in the set of interior-vertices in G.
9. $\mathrm{dcp}_i(G)$, $i = 17 + |\Lambda^{\mathrm{int}}(D_\pi)| + [\mathtt{a}]^{\mathrm{ex}}$, $\mathtt{a} \in \Lambda^{\mathrm{ex}}(D_\pi)$: the frequency $\mathrm{na}_{\mathtt{a}}^{\mathrm{ex}}(G)$ of chemical element $\mathtt{a}$ in the set of exterior-vertices in G.
10. $\mathrm{dcp}_i(G)$, $i = 17 + |\Lambda^{\mathrm{int}}(D_\pi)| + |\Lambda^{\mathrm{ex}}(D_\pi)| + [\gamma]$, $\gamma \in \Gamma^{\mathrm{int}}(D_\pi)$: the frequency $\mathrm{ec}_\gamma(G)$ of edge-configuration γ in the set of interior-edges $e \in E^{\mathrm{int}}$ in G.
11. $\mathrm{dcp}_i(G)$, $i = 17 + |\Lambda^{\mathrm{int}}(D_\pi)| + |\Lambda^{\mathrm{ex}}(D_\pi)| + |\Gamma^{\mathrm{int}}(D_\pi)| + [\psi]$, $\psi \in \mathcal{F}(D_\pi)$: the frequency $\mathrm{fc}_\psi(G)$ of fringe-configuration ψ in the set of ρ-fringe-trees in G.

2.2. Specifying Target Chemical Graphs

Given a prediction function $\eta_{\mathcal{N}}$ and a target value $y^* \in \mathbb{R}$, we call a chemical graph G^* such that $\eta_{\mathcal{N}}(x^*) = y^*$ for the feature vector $x^* = f(G^*)$ a *target chemical graph*. This section presents a set of rules for specifying topological substructure of a target chemical graph in a flexible way in Stage 4.

We first describe how to reduce a chemical graph $G = (H, \alpha, \beta)$ into an abstract form based on which our specification rules will be defined. To illustrate the reduction process, we use the chemical graph $G = (H, \alpha, \beta)$ in Figure 2.

R1 **Removal of all ρ-fringe-trees:** The interior $H^{\mathrm{int}} = (V^{\mathrm{int}}(H), E^{\mathrm{int}}(H))$ of G is obtained by removing the non-root vertices of each ρ-fringe-trees $T \in \mathcal{T}^{\mathrm{ex}}(G)$. Figure 4 illustrates the interior H^{int} of chemical graph G with $\rho = 2$ in Figure 2.

R2 **Removal of some leaf paths:** We call a u, v-path Q in H^{int} a *leaf path* if vertex v is a leaf-vertex of H^{int} and the degree of each internal vertex of Q in H^{int} is 2, where we regard that Q is rooted at vertex u. A connected subgraph S of the interior H^{int} of G is called a *cyclical-base* if S is obtained from H by removing the vertices in $V(Q_u) \setminus \{u\}, u \in X$ for a subset X of interior-vertices and a set $\{Q_u \mid u \in X\}$ of leaf u, v-paths Q_u such that no two paths Q_u and $Q_{u'}$ share a vertex. Figure 5a illustrates a cyclical-base $S = H^{\mathrm{int}} - \bigcup_{u \in X}(V(Q_u) \setminus \{u\})$ of the interior H^{int} for a set $\{Q_{u_5} = (u_5, u_{24}), Q_{u_{18}} = (u_{18}, u_{25}, u_{26}, u_{27}), Q_{u_{22}} = (u_{22}, u_{28})\}$ of leaf paths in Figure 4.

R3 **Contraction of some pure paths:** A path in S is called *pure* if each internal vertex of the path is of degree 2. Choose a set $\mathcal{P}$ of several pure paths in S so that no two paths share vertices except for their end-vertices. A graph S' is called a *contraction* of a graph S (with respect to $\mathcal{P}$) if S' is obtained from S by replacing each pure u, v-path with a single edge $a = uv$, where S' may contain multiple edges between the same pair of adjacent vertices. Figure 5b illustrates a contraction S' obtained from the chemical graph S by contracting each uv-path $P_a \in \mathcal{P}$ into a new edge $a = uv$, where $a_1 = u_1u_2, a_2 = u_1u_3, a_3 = u_4u_7, a_4 = u_{10}u_{11}$, and $a_5 = u_{11}u_{12}$, and $\mathcal{P} = \{P_{a_1} = (u_1, u_{13}, u_2), P_{a_2} = (u_1, u_{14}, u_3), P_{a_3} = (u_4, u_{15}, u_{16}, u_7), P_{a_4} = (u_{10}, u_{17}, u_{18}, u_{19}, u_{11}), P_{a_5} = (u_{11}, u_{20}, u_{21}, u_{22}, u_{12})\}$ of pure paths in Figure 5a.

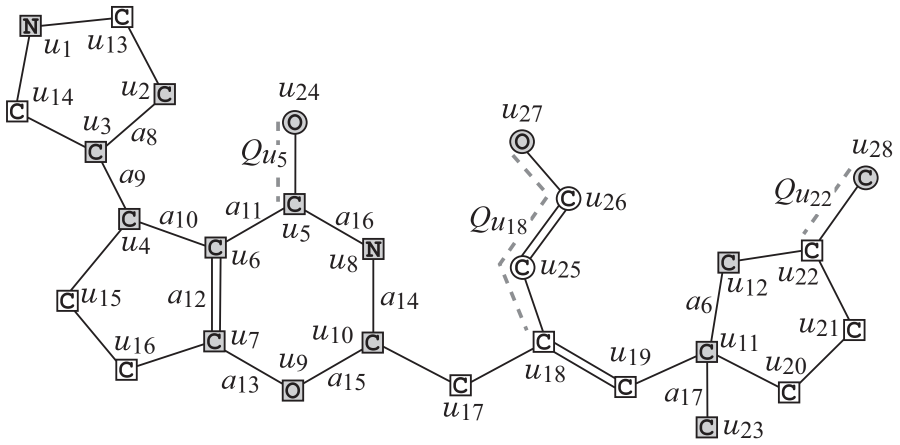
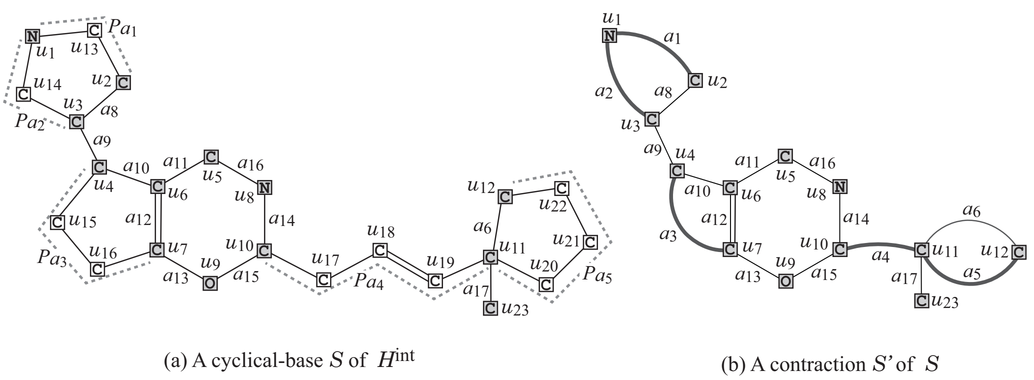


Figure 4. The interior H^{int} of chemical graph G with $\rho = 2$ in Figure 2.



(a) A cyclical-base S of H^{int}

(b) A contraction S' of S

Figure 5. (**a**) A cyclical-base $S = H^{\text{int}} - \bigcup_{u \in \{u_5, u_{18}, u_{22}\}}(V(Q_u) \setminus \{u\})$ of the interior H^{int} in Figure 4; (**b**) A contraction S' of S for a pure path set $\mathcal{P} = \{P_{a_1}, P_{a_2}, \ldots, P_{a_5}\}$ in (**a**), where a new edge obtained by contracting a pure path is depicted with a thick line.

We will define a set of rules so that a chemical graph can be obtained from a graph (called a seed graph in the next section) by applying processes R3 to R1 in a reverse way. We specify topological substructures of a target chemical graph with a tuple $(G_{\text{C}}, \sigma_{\text{int}}, \sigma_{\text{ce}})$ called a *target specification* defined under the set of the following rules.

Seed Graphs

A *seed graph* $G_{\text{C}} = (V_{\text{C}}, E_{\text{C}})$ is defined to be a graph (possibly with multiple edges) such that the edge set E_{C} consists of four sets $E_{(\geq 2)}$, $E_{(\geq 1)}$, $E_{(0/1)}$, and $E_{(=1)}$, where each of them can be empty. A seed graph plays a role of the most abstract form S' in R3. Figure 3a illustrates an example of a seed graph, where $V_{\text{C}} = \{u_1, u_2, \ldots, u_{12}\}$, $E_{(\geq 2)} = \{a_1, a_2, \ldots, a_5\}$, $E_{(\geq 1)} = \{a_6\}$, $E_{(0/1)} = \{a_7\}$, and $E_{(=1)} = \{a_8, a_9, \ldots, a_{17}\}$.

A *subdivision* S of G_{C} is a graph constructed from a seed graph G_{C} according to the following rules:

- Each edge $e = uv \in E_{(\geq 2)}$ is replaced with a u,v-path P_e of length at least 2;
- Each edge $e = uv \in E_{(\geq 1)}$ is replaced with a u,v-path P_e of length at least 1 (equivalently e is directly used or replaced with a u,v-path P_e of length at least 2);
- Each edge $e \in E_{(0/1)}$ is either used or discarded; and
- Each edge $e \in E_{(=1)}$ is always used directly.

We allow a possible elimination of edges in $E_{(0/1)}$ as an optional rule in constructing a target chemical graph from a seed graph, even though such an operation has not been included in the process R3. A subdivision S plays a role of a cyclical-base in R2. A target chemical graph $G = (H, \alpha, \beta)$ will contain S as a subgraph of the interior $H^{\rm int}$ of G.

Interior-Specification

A graph H^* that serves as the interior $H^{\rm int}$ of a target chemical graph G will be constructed as follows. First, construct a subdivision S of a seed graph $G_{\rm C}$ by replacing each edge edge $e = uu' \in E_{(\geq 2)} \cup E_{(\geq 1)}$ with a pure u, u'-path P_e. Next, construct a supergraph H^* of S by attaching a leaf path Q_v at each vertex $v \in V_{\rm C}$ or at an internal vertex $v \in V(P_e) \setminus \{u, u'\}$ of each pure u, u'-path P_e for some edge $e = uu' \in E_{(\geq 2)} \cup E_{(\geq 1)}$, where possibly $Q_v = v, E(Q_v) = \emptyset$ (i.e., we do not attach any new edges to v). We introduce the following rules for specifying the size of H^*, the length $|E(P_e)|$ of a pure path P_e, the length $|E(Q_v)|$ of a leaf path Q_v, the number of leaf paths Q_v, and a bond-multiplicity of each interior-edge, where we call the set of prescribed constants an *interior-specification* $\sigma_{\rm int}$:

- Lower and upper bounds $\mathrm{n}^{\rm int}_{\rm LB}, \mathrm{n}^{\rm int}_{\rm UB} \in \mathbb{Z}_+$ on the number of interior-vertices of a target chemical graph G.
- For each edge $e = uu' \in E_{(\geq 2)} \cup E_{(\geq 1)}$,
  - a lower bound $\ell_{\rm LB}(e)$ and an upper bound $\ell_{\rm UB}(e)$ on the length $|E(P_e)|$ of a pure u, u'-path P_e. (For a notational convenience, set $\ell_{\rm LB}(e) := 0$, $\ell_{\rm UB}(e) := 1$, $e \in E_{(0/1)}$ and $\ell_{\rm LB}(e) := 1$, $\ell_{\rm UB}(e) := 1$, $e \in E_{(=1)}$.)
  - a lower bound $\mathrm{bl}_{\rm LB}(e)$ and an upper bound $\mathrm{bl}_{\rm UB}(e)$ on the number of leaf paths Q_v attached to at internal vertices v of a pure u, u'-path P_e.
  - a lower bound $\mathrm{ch}_{\rm LB}(e)$ and an upper bound $\mathrm{ch}_{\rm UB}(e)$ on the maximum length $|E(Q_v)|$ of a leaf path Q_v attached at an internal vertex $v \in V(P_e) \setminus \{u, u'\}$ of a pure u, u'-path P_e.
- For each vertex $v \in V_{\rm C}$,
  - a lower bound $\mathrm{ch}_{\rm LB}(e)$ and an upper bound $\mathrm{ch}_{\rm UB}(e)$ on the number of leaf paths Q_v attached to v, where $0 \leq \mathrm{ch}_{\rm LB}(e) \leq \mathrm{ch}_{\rm UB}(e) \leq 1$.
  - a lower bound $\mathrm{ch}_{\rm LB}(v)$ and an upper bound $\mathrm{ch}_{\rm UB}(v)$ on the length $|E(Q_v)|$ of a leaf path Q_v attached to v.
- For each edge $e = uu' \in E_{\rm C}$, a lower bound $\mathrm{bd}_{m,\rm LB}(e)$ and an upper bound $\mathrm{bd}_{m,\rm UB}(e)$ on the number of edges with bond-multiplicity $m \in [2, 3]$ in u, u'-path P_e, where we regard P_e, $e \in E_{(0/1)} \cup E_{(=1)}$ as single edge e.

We call a graph H^* that satisfies an interior-specification $\sigma_{\rm int}$ a $\sigma_{\rm int}$*-extension of* $G_{\rm C}$, where the bond-multiplicity of each edge has been determined.

Table 1 shows an example of an interior-specification $\sigma_{\rm int}$ to the seed graph $G_{\rm C}$ in Figure 3.

Figure 6 illustrates an example of an $\sigma_{\rm int}$-extension H^* of seed graph $G_{\rm C}$ in Figure 3 under the interior-specification $\sigma_{\rm int}$ in Table 1.

Chemical-Specification

Let H^* be a graph that serves as the interior $H^{\rm int}$ of a target chemical graph G, where the bond-multiplicity of each edge in H^* has be determined. Finally, we introduce a set of rules for constructing a target chemical graph G from H^* by choosing a chemical element $\mathtt{a} \in \Lambda$ and assigning a ρ-fringe-tree ψ to each interior-vertex $v \in V^{\rm int}$. We introduce the following rules for specifying the size of G, a set of chemical rooted trees that are allowed to use as ρ-fringe-trees and lower and upper bounds on the frequency of a chemical element, a chemical symbol, and an edge-configuration, where we call the set of prescribed constants a *chemical specification* $\sigma_{\rm ce}$:

- Lower and upper bounds $n_{\rm LB}, n^* \in \mathbb{Z}_+$ on the number of vertices in G, where $\mathrm{n}^{\rm int}_{\rm LB} \leq n_{\rm LB} \leq n^*$.
- Subsets $\mathcal{F}(v) \subseteq \mathcal{F}(D_\pi), v \in V_{\rm C}$ and $\mathcal{F}_E \subseteq \mathcal{F}(D_\pi)$ of chemical rooted trees with height at most ρ, where we require that every ρ-fringe-tree T_v rooted at a vertex $v \in V_{\rm C}$ (resp., at an internal vertex v not in $V_{\rm C}$) in G belongs to $\mathcal{F}(v)$ (resp., $\mathcal{F}_E$). Let

$\mathcal{F}^* := \mathcal{F}_E \cup \bigcup_{v \in V_C} \mathcal{F}(v)$ and $\Lambda^{\rm ex}$ denote the set of chemical elements assigned to non-root vertices over all chemical rooted trees in $\mathcal{F}^*$.

- A subset $\Lambda^{\rm int} \subseteq \Lambda^{\rm int}(D_\pi)$, where we require that every chemical element $\alpha(v)$ assigned to an interior-vertex v in G belongs to $\Lambda^{\rm int}$. Let $\Lambda := \Lambda^{\rm int} \cup \Lambda^{\rm ex}$ and $\mathrm{na}_{\mathtt{a}}(G)$ (resp., $\mathrm{na}^{\rm int}_{\mathtt{a}}(G)$ and $\mathrm{na}^{\rm ex}_{\mathtt{a}}(G)$) denote the number of vertices (resp., interior-vertices and exterior-vertices) v such that $\alpha(v) = \mathtt{a}$ in G.
- A set $\Lambda^{\rm int}_{\rm dg} \subseteq \Lambda \times [1,4]$ of chemical symbols and a set $\Gamma^{\rm int} \subseteq \Gamma^{\rm int}(D_\pi)$ of edge-configurations (μ, ξ, m) with $\mu \leq \xi$, where we require that the edge-configuration $\mathrm{ec}(e)$ of an interior-edge e in G belongs to $\Gamma^{\rm int}$. We do not distinguish (μ, ξ, m) and (ξ, μ, m).
- Define $\Gamma^{\rm int}_{\rm ac}$ to be the set of adjacency-configurations such that $\Gamma^{\rm int}_{\rm ac} := \{(\mathtt{a}, \mathtt{b}, m) \mid (\mathtt{a}d, \mathtt{b}d', m) \in \Gamma^{\rm int}\}$. Let $\mathrm{ac}^{\rm int}_\nu(G), \nu \in \Gamma^{\rm int}_{\rm ac}$ denote the number of interior-edges e such that $\mathrm{ac}(e) = \nu$ in G.
- Subsets $\Lambda^*(v) \subseteq \{\mathtt{a} \in \Lambda^{\rm int} \mid \mathrm{val}(\mathtt{a}) \geq 2\}$, $v \in V_{\rm C}$, we require that every chemical element $\alpha(v)$ assigned to a vertex $v \in V_{\rm C}$ in the seed graph belongs to $\Lambda^*(v)$.
- Lower and upper bound functions $\mathrm{na}_{\rm LB}, \mathrm{na}_{\rm UB} : \Lambda \to [1, n^*]$ and $\mathrm{na}^{\rm int}_{\rm LB}, \mathrm{na}^{\rm int}_{\rm UB} : \Lambda^{\rm t} \to [1, n^*]$ on the number of interior-vertices v such that $\alpha(v) = \mathtt{a}$ in G.
- Lower and upper bound functions $\mathrm{ns}^{\rm int}_{\rm LB}, \mathrm{ns}^{\rm int}_{\rm UB} : \Lambda^{\rm int}_{\rm dg} \to [1, n^*]$ on the number of interior-vertices v such that $\mathrm{cs}(v) = \mu$ in G.
- Lower and upper bound functions $\mathrm{ac}^{\rm int}_{\rm LB}, \mathrm{ac}^{\rm int}_{\rm UB} : \Gamma^{\rm int}_{\rm ac} \to \mathbb{Z}_+$ on the number of interior-edges e such that $\mathrm{ac}(e) = \nu$ in G.
- Lower and upper bound functions $\mathrm{ec}^{\rm int}_{\rm LB}, \mathrm{ec}^{\rm int}_{\rm UB} : \Gamma^{\rm int} \to \mathbb{Z}_+$ on the number of interior-edges e such that $\mathrm{ec}(e) = \gamma$ in G.

Table 1. Example 1 of an interior-specification $\sigma_{\rm int}$.

$\mathbf{n^{int}_{LB} = 20}$	$\mathbf{n^{int}_{UB} = 28}$

	a_1	a_2	a_3	a_4	a_5	a_6
$\ell_{\rm LB}(a_i)$	2	2	2	3	2	1
$\ell_{\rm UB}(a_i)$	3	4	3	5	4	4
$\mathrm{bl}_{\rm LB}(a_i)$	0	0	0	1	1	0
$\mathrm{bl}_{\rm UB}(a_i)$	1	1	0	2	1	0
$\mathrm{ch}_{\rm LB}(a_i)$	0	1	0	4	3	0
$\mathrm{ch}_{\rm UB}(a_i)$	3	3	1	6	5	2

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	u_{11}	u_{12}	u_{23}
$\mathrm{bl}_{\rm LB}(u_i)$	0	0	0	0	0	0	0	0	0	0	0	0	0
$\mathrm{bl}_{\rm UB}(u_i)$	1	1	1	1	1	0	0	0	0	0	0	0	0
$\mathrm{ch}_{\rm LB}(u_i)$	0	0	0	0	1	0	0	0	0	0	0	0	0
$\mathrm{ch}_{\rm UB}(u_i)$	1	0	0	0	3	0	1	1	0	1	2	4	1

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{17}
$\mathrm{bd}_{2,\rm LB}(a_i)$	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
$\mathrm{bd}_{2,\rm UB}(a_i)$	1	1	0	2	2	0	0	0	0	0	0	1	0	0	0	0	0
$\mathrm{bd}_{3,\rm LB}(a_i)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$\mathrm{bd}_{3,\rm UB}(a_i)$	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

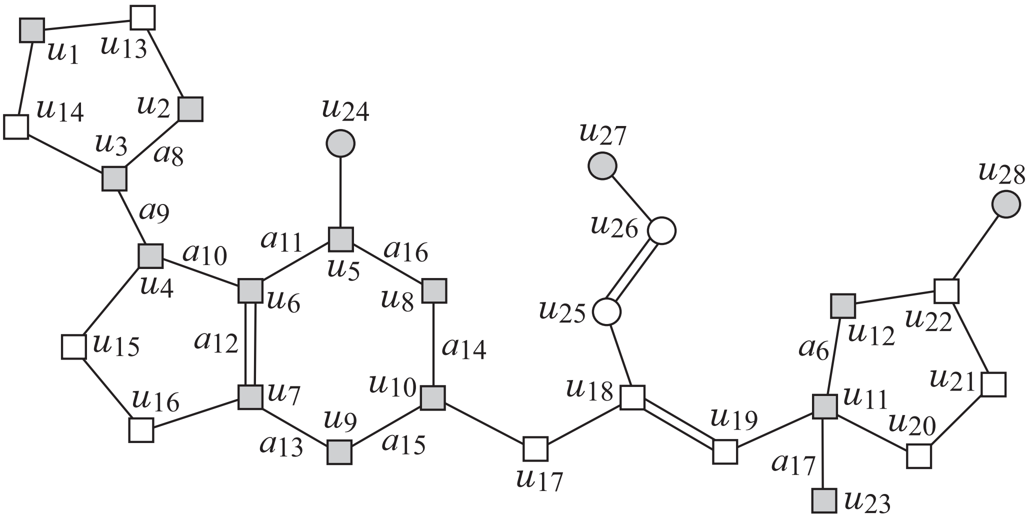


Figure 6. An illustration of a graph H^* that is obtained from the seed graph G_{C} in Figure 3 under the interior-specification σ_{int} in Table 1, where the vertices newly introduced by pure paths P_{a_i} and leaf paths Q_{v_i} are depicted with white squares and circles, respectively.

We call a chemical graph G that satisfies a chemical specification σ_{ce} a $(\sigma_{\mathrm{int}}, \sigma_{\mathrm{ce}})$-*extension of* G_{C}, and denote by $\mathcal{G}(G_{\mathrm{C}}, \sigma_{\mathrm{int}}, \sigma_{\mathrm{ce}})$ the set of all $(\sigma_{\mathrm{int}}, \sigma_{\mathrm{ce}})$-extensions of G_{C}.

Table 2 shows an example of a chemical-specification σ_{ce} to the seed graph G_{C} in Figure 3.

Table 2. Example 2 of a chemical-specification σ_{ce}.

$\boldsymbol{n_{\mathrm{LB}} = 30, n^* = 50}$.
branch-parameter: $\rho = 2$
Each of sets $\mathcal{F}(v), v \in V_{\mathrm{C}}$ and $\mathcal{F}_E$ is set to be the set $\mathcal{F}$ of chemical rooted trees with height at most $\rho = 2$ in Figure 3b.
$\Lambda = \{\mathtt{C}, \mathtt{N}, \mathtt{O}\}$ $\Lambda_{\mathrm{dg}} = \{\mathtt{C2}, \mathtt{C3}, \mathtt{C4}, \mathtt{N2}, \mathtt{O2}\}$
$\Gamma^{\mathrm{int}}_{\mathrm{ac}}$ $\nu_1 = (\mathtt{C}, \mathtt{C}, 1), \nu_2 = (\mathtt{C}, \mathtt{C}, 2), \nu_3 = (\mathtt{C}, \mathtt{N}, 1), \nu_4 = (\mathtt{C}, \mathtt{O}, 1)$
Γ^{int} $\gamma_1 = (\mathtt{C2}, \mathtt{C2}, 1), \gamma_2 = (\mathtt{C2}, \mathtt{C3}, 1), \gamma_3 = (\mathtt{C2}, \mathtt{C3}, 2), \gamma_4 = (\mathtt{C2}, \mathtt{C4}, 1), \gamma_5 = (\mathtt{C3}, \mathtt{C3}, 1)$, $\gamma_6 = (\mathtt{C3}, \mathtt{C3}, 2), \gamma_7 = (\mathtt{C3}, \mathtt{C4}, 1), \gamma_8 = (\mathtt{C2}, \mathtt{N2}, 1), \gamma_9 = (\mathtt{C3}, \mathtt{N2}, 1), \gamma_{10} = (\mathtt{C3}, \mathtt{O2}, 1)$, $\gamma_{11} = (\mathtt{C2}, \mathtt{C2}, 2), \gamma_{12} = (\mathtt{C2}, \mathtt{O2}, 1)$,
$\Lambda^*(u_1) = \{\mathtt{N}\}, \Lambda^*(u_8) = \{\mathtt{C}, \mathtt{N}\}, \Lambda^*(u_9) = \{\mathtt{C}, \mathtt{O}\}, \Lambda^*(u) = \{\mathtt{C}\}, u \in V_{\mathrm{C}} \setminus \{u_1, u_8, u_9\}$

	$\mathtt{C}$	$\mathtt{N}$	$\mathtt{O}$
$\mathrm{na}_{\mathrm{LB}}(\mathtt{a})$	27	1	1
$\mathrm{na}_{\mathrm{UB}}(\mathtt{a})$	37	4	8

	$\mathtt{C}$	$\mathtt{N}$	$\mathtt{O}$
$\mathrm{na}^{\mathrm{int}}_{\mathrm{LB}}(\mathtt{a})$	9	1	0
$\mathrm{na}^{\mathrm{int}}_{\mathrm{UB}}(\mathtt{a})$	23	4	5

	$\mathtt{C2}$	$\mathtt{C3}$	$\mathtt{C4}$	$\mathtt{N2}$	$\mathtt{N3}$	$\mathtt{O2}$
$\mathrm{ns}^{\mathrm{int}}_{\mathrm{LB}}(\mu)$	3	5	0	0	0	0
$\mathrm{ns}^{\mathrm{int}}_{\mathrm{UB}}(\mu)$	8	15	2	2	3	5

	ν_1	ν_2	ν_3	ν_4
$\mathrm{ac}^{\mathrm{int}}_{\mathrm{LB}}(\nu)$	0	0	0	0
$\mathrm{ac}^{\mathrm{int}}_{\mathrm{UB}}(\nu)$	30	10	10	10

	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8	γ_9	γ_{10}	γ_{11}	γ_{12}
$\mathrm{ec}^{\mathrm{int}}_{\mathrm{LB}}(\gamma)$	0	0	0	0	0	0	0	0	0	0	0	0
$\mathrm{ec}^{\mathrm{int}}_{\mathrm{UB}}(\gamma)$	4	15	4	4	10	5	4	4	6	4	4	4

Figure 2 illustrates an example of a $(\sigma_{\mathrm{int}}, \sigma_{\mathrm{ce}})$-extension of G_{C} obtained from the

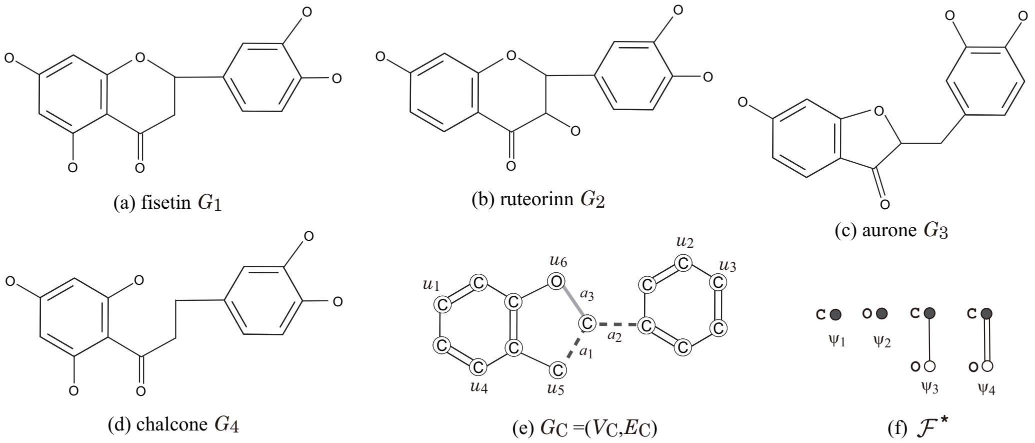
$\sigma_{\rm int}$-extension H^* in Figure 6 under the chemical-specification $\sigma_{\rm ce}$ in Table 2.

Our specification of topological substructures is similar to that proposed by Akutsu and Nagamochi [27], wherein a target chemical graph is restricted to ρ-lean cyclic graphs and prescribed substructures cannot be specified in the acyclic part. In our new method, a chemical graph with any structure can be handled and substructures in the acyclic part can be fixed.

2.3. Examples of Specification

We here present some cases where a target specification $(G_{\rm C}, \sigma_{\rm int}, \sigma_{\rm ce})$ can be chosen based on a set $\mathcal{G}^*$ of given chemical graphs with a similar structure so that $\mathcal{G}^*$ becomes a subset of $\mathcal{G}(G_{\rm C}, \sigma_{\rm int}, \sigma_{\rm ce})$. In such a case, every target chemical graph in $\mathcal{G}(G_{\rm C}, \sigma_{\rm int}, \sigma_{\rm ce})$ possesses a common structure over the given set $\mathcal{G}^*$.

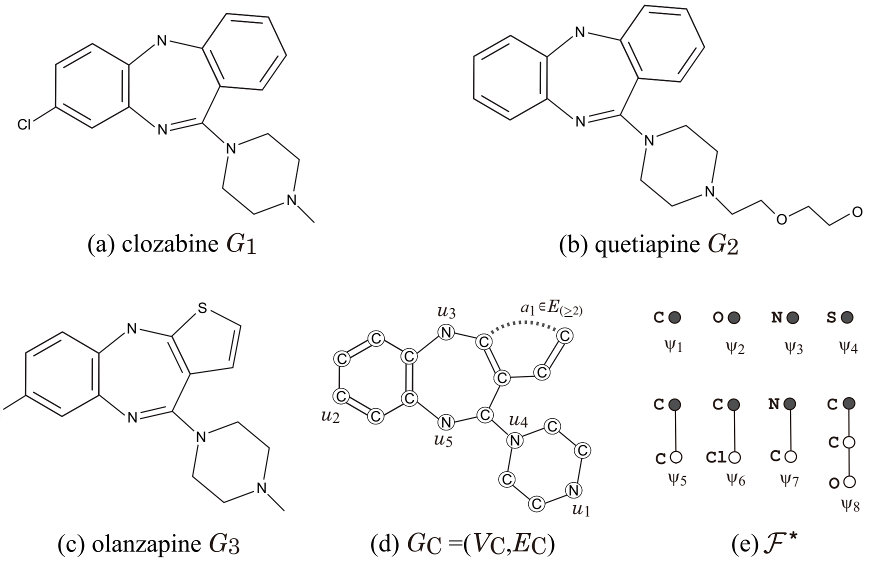
Figure 7 illustrates a set $\mathcal{G}^*$ of four flavonoids and a seed graph $G_{\rm C}$ for $\rho = 2$ so that $\mathcal{G}^* \subseteq \mathcal{G}(G_{\rm C}, \sigma_{\rm int}, \sigma_{\rm ce})$ for a choice of an interior-specification $\sigma_{\rm int}$ and a chemical-specification $\sigma_{\rm ce}$. Let $\Lambda := \{\texttt{C}, \texttt{O}\}$. In the seed graph $G_{\rm C} = (V_{\rm C}, E_{\rm C})$, we set $E_{(\geq 1)} := \{a_1, a_2\}$, $E_{(0/1)} := \{a_3\}$, and $E_{(=1)} := E_{\rm C} \setminus \{a_1, a_2, a_3\}$ and predetermine the chemical element $\alpha(u)$ for each vertex $u \in V_{\rm C}$ and the bond-multiplicity $\beta(e)$ for each edge $e \in E_{(=1)}$ as in Figure 7e, i.e., $\Lambda^*(u) := \{\texttt{a}\}$ for $\texttt{a} = \alpha(u)$ and $\mathrm{bd}_{m,\rm LB}(e) := 1$ for $m = \beta(e)$. Figure 7f illustrates a set $\mathcal{F}^*$ of chemical rooted trees for the 2-fringe-trees in a target chemical graph. For vertices in $G_{\rm C}$, we set $\mathrm{ch}_{\rm UB}(u) := 0, u \in V_{\rm C}$, $\mathcal{F}(u_i) := \{\psi_3\}, i \in [1,3]$, $\mathcal{F}(u_4) := \{\psi_1, \psi_3\}$, $\mathcal{F}(u_5) := \{\psi_4\}$, $\mathcal{F}(u_6) := \{\psi_2\}$, and $\mathcal{F}(u) := \{\psi_1\}, u \in V_{\rm C} \setminus \{u_1, u_2, \ldots, u_6\}$. For edges $a_i \in E_{(\geq 1)}, i = 1, 2$, we set $\ell_{\rm UB}(a_i) := 2, \mathrm{ch}_{\rm UB}(a_i) := 0$ and $\mathcal{F}_E := \{\psi_1, \psi_2\}$, where a pure path P_{a_i} may be introduced in a target chemical graph. We see that every given chemical graph $G_i \in \mathcal{G}^*$ belongs to $\mathcal{G}(G_{\rm C}, \sigma_{\rm int}, \sigma_{\rm ce})$ by setting the other specification in $\sigma_{\rm int}$ and $\sigma_{\rm ce}$ adequately.



(a) fisetin G_1 (b) ruteorinn G_2 (c) aurone G_3 (d) chalcone G_4 (e) $G_{\rm C} = (V_{\rm C}, E_{\rm C})$ (f) $\mathcal{F}^*$

Figure 7. Illustration of a set $\mathcal{G}^* = \{G_1, G_2, G_3, G_4\}$ of four flavonoids, a seed graph $G_{\rm C}$, and a set $\mathcal{F}^* = \{\psi_1, \psi_2, \psi_3, \psi_4\}$ of chemical rooted trees for $\rho = 2$: (**a**) fisetin G_1; (**b**) ruteorinn G_2; (**c**) aurone G_3; (**d**) chalcone G_4; (**e**) $G_{\rm C} = (V_{\rm C}, E_{\rm C})$; (**f**) $\mathcal{F}^* = \mathcal{F}_E \cup \bigcup_{v \in V_{\rm C}} \mathcal{F}(v)$.

Figure 8 illustrates a set $\mathcal{G}^*$ of three dibenzodiazepine atypical antipsychotics, and a seed graph $G_{\rm C}$ for $\rho = 2$ so that $\mathcal{G}^* \subseteq \mathcal{G}(G_{\rm C}, \sigma_{\rm int}, \sigma_{\rm ce})$ for a choice of an interior-specification $\sigma_{\rm int}$ and a chemical-specification $\sigma_{\rm ce}$. Let $\Lambda := \{\texttt{C}, \texttt{O}, \texttt{N}, \texttt{S}, \texttt{Cl}\}$. In the seed graph $G_{\rm C} = (V_{\rm C}, E_{\rm C})$, we set $E_{(\geq 2)} := \{a_1\}$ and $E_{(=1)} := E_{\rm C} \setminus \{a_1\}$ and predetermine the chemical element $\alpha(u)$ for each vertex $u \in V_{\rm C}$ and the bond-multiplicity $\beta(e)$ for each edge $e \in E_{(=1)}$ as in Figure 8d. Figure 8e illustrates a set $\mathcal{F}^*$ of chemical rooted trees for the 2-fringe-trees in a target chemical graph. For vertices in $G_{\rm C}$, we set $\mathrm{ch}_{\rm UB}(u) := 0, u \in V_{\rm C} \setminus \{u_2\}$, $\mathrm{ch}_{\rm LB}(u_2) := 0$, $\mathrm{ch}_{\rm UB}(u_2) := 4$, $\mathcal{F}(u_1) := \{\psi_3, \psi_7\}$, $\mathcal{F}(u_2) := \{\psi_1, \psi_6\}$, $\mathcal{F}(u_i) := \{\psi_3\}, i \in [3,5]$, and $\mathcal{F}(u) := \{\psi_1\}, u \in V_{\rm C} \setminus \{u_1, u_2, \ldots, u_6\}$, where a leaf path Q_{u_2} may be introduced in a target chemical graph. For edge $a_1 \in E_{(\geq 2)}$, we set $\ell_{\rm UB}(a_1) := 3, \mathrm{ch}_{\rm UB}(a_1) := 0$ and $\mathcal{F}_E := \{\psi_1, \psi_2, \psi_4, \psi_8\}$. We see that every given chemical graph $G_i \in \mathcal{G}^*$ belongs to $\mathcal{G}(G_{\rm C}, \sigma_{\rm int}, \sigma_{\rm ce})$ by setting the other specification in $\sigma_{\rm int}$ and $\sigma_{\rm ce}$ adequately.



(a) clozabine G_1 (b) quetiapine G_2 (c) olanzapine G_3 (d) $G_{\rm C} = (V_{\rm C}, E_{\rm C})$ (e) $\mathcal{F}^\star$

Figure 8. Illustration of a set $\mathcal{G}^* = \{G_1, G_2, G_3\}$ of three dibenzodiazepine atypical antipsychotics, a seed graph $G_{\rm C}$ and a set $\mathcal{F}^* = \{\psi_1, \psi_2, \ldots, \psi_8\}$ of chemical rooted trees for $\rho = 2$: (**a**) clozabine G_1; (**b**) quetiapine G_2; (**c**) olanzapine G_3; (**d**) $G_{\rm C} = (V_{\rm C}, E_{\rm C})$; (**e**) $\mathcal{F}^* = \mathcal{F}_E \cup \bigcup_{v \in V_{\rm C}} \mathcal{F}(v)$.

3. Results

We implemented our method of Stages 1 to 5 for inferring chemical graphs under a given target specification and conducted experiments to evaluate the computational efficiency. We executed the experiments on a PC with Processor: 3.0 GHz Core i7-9700 (3.0 GHz) Memory: 16 GB RAM DDR4. We used ChemDoodle version 10.2.0 for constructing 2D drawings of chemical graphs.

To conduct experiments for Stages 1 to 5, we selected six chemical properties π: octanol/water partition coefficient (Kow), boiling point (Bp), melting point (Mp), flash point (closed cup) (Fp), lipophylicity (Lp), solubility (Sl) provided by HSDB from PubChem [29] for Kow, Bp, Mp, and Fp, figshare [30] for Lp and MoleculeNet [31] for Sl.

Results on Phase 1.

We implemented Stages 1, 2, and 3 in Phase 1 as follows.

Stage 1. We set a graph class $\mathcal{G}$ to be the set of all chemical graphs with any graph structure, and set a branch-parameter ρ to be 2. For each property $\pi \in \{$Kow, Bp, Mp, Fp, Lp, Sl$\}$, we first select a set Λ of chemical elements and then collect a data set D_π on chemical graphs over the set Λ of chemical elements. To construct the data set D_π, we eliminated chemical compounds that have at most three carbon atoms or contain a charged element such as $\mathtt{N}^+$ or an element $\mathtt{a} \in \Lambda$ whose valence is different from our setting of valence function val.

Table 3 shows the size and range of data sets that we prepared for each chemical property in Stage 1, where we denote the following:

Λ: the set of selected chemical elements (hydrogen atoms are added at the final stage);
$|D_\pi|$: the size of data set D_π over Λ for property π;
$|\Gamma^{\rm int}(D_\pi)|$: the number of different edge-configurations of interior-edges over the compounds in D_π;
$|\mathcal{F}(D_\pi)|$: the number of non-isomorphic chemical rooted trees in the set of all 2-fringe-trees in the compounds in D_π;
$[\underline{n}, \overline{n}]$: the minimum and maximum values of $n(G)$ over the compounds G in D_π; and
$[\underline{a}, \overline{a}]$: the minimum and maximum values of $a(G)$ in π over compounds G in D_π.

Table 3. Data sets for stage 1 in phase 1.

π	Λ	$\lvert D_\pi \rvert$	$\lvert \Gamma^{\mathrm{int}}(D_\pi) \rvert$	$\lvert \mathcal{F}(D_\pi) \rvert$	$[\underline{n}, \overline{n}]$	$[\underline{a}, \overline{a}]$
Kow	`C,O,N`	644	24	109	[4, 58]	[−7.53, 13.45]
Kow	`C,O,N,S,Cl`	837	31	142	[4, 73]	[−7.53, 13.45]
Bp	`C,O,N`	358	21	91	[4, 30]	[−11.70, 470.0]
Bp	`C,O,N,S,Cl`	425	23	114	[4, 30]	[−11.70, 470.0]
Mp	`C,O,N`	448	22	94	[4, 122]	[−185.3, 300.0]
Mp	`C,O,N,S,Cl`	548	26	118	[4, 122]	[−185.3, 300.0]
Fp	`C,O,N`	348	20	85	[4, 66]	[−82.99, 300.0]
Fp	`C,O,N,S,Cl`	399	24	107	[4, 66]	[−82.99, 300.0]
Lp	`C,O,N`	592	27	71	[6, 60]	[−3.62, 6.84]
Lp	`C,O,N,S,Cl`	779	32	78	[6, 74]	[−3.62, 6.84]
Sl	`C,O,N`	640	25	111	[4, 55]	[−9.33, 1.11]
Sl	`C,O,N,S,Cl`	847	31	144	[4, 55]	[−11.60, 1.11]

Stage 2. We used the new feature function that consists of the descriptors such as fringe-configuration defined in Section 2.1 and let f_{fc} denote the feature function.

Stage 3. Let $\eta : \mathbb{R}^K \to \mathbb{R}$ be a prediction function to a property function $a : D \to \mathbb{R}$ with a feature function $f : D \to \mathbb{R}^K$ for a data set D of chemical graphs. We define the coefficient of determination $\mathrm{R}^2(f, \eta, D)$ of a prediction function η over a data set D to be

$$\mathrm{R}^2(f, \eta, D) \triangleq 1 - \frac{\sum_{G \in D}(a(G) - \eta(f(G)))^2}{\sum_{G \in D}(a(G) - \widetilde{a})^2} \text{ for } \widetilde{a} = \frac{1}{|D|}\sum_{G \in D} a(G).$$

To conduct an experiment in Stage 3, we first constructed ten architectures A_j, $j \in [1, 10]$ with one or two hidden layers. For each pair (π, A_j) of a property $\pi \in \{$Kow, Bp, Mp, Fp, Lp, Sl$\}$, and an architecture A_j, $j \in [1, 10]$, we constructed five prediction functions in order to evaluate the performance with cross-validation as follows. Partition data set D_π into five subsets $D_\pi^{(i)}$, $i \in [1, 5]$ randomly and for each set $D_\pi \setminus D_\pi^{(i)}$ construct an ANN $\mathcal{N}(j, i)$ and its prediction function $\eta_{\mathcal{N}(j,i)}$ using the feature function f_{fc}. We used `scikit-learn` version 0.23.2 with Python 3.8.5, MLPRegressor and ReLU activation function to construct each ANN $\mathcal{N}(j, i)$. We evaluated the resulting prediction function $\eta_{\mathcal{N}(j,i)}$ with the coefficient $\mathrm{R}^2(f_{\mathrm{fc}}, \eta_{\mathcal{N}(j,i)}, D_\pi^{(i)})$ of determination for the test set $D_\pi^{(i)}$. For each property π, let t-$\mathrm{R}^2_{\mathrm{cv}}(j)$ denote the average of $\mathrm{R}^2(f_{\mathrm{fc}}, \eta_{\mathcal{N}(j,i)}, D_\pi^{(i)})$ over all $i \in [1, 5]$ in the cross-validation to an architecture A_j.

Table 4 shows the results on Stages 2 and 3, where we denote the following.

- Λ: the set of selected chemical elements (hydrogen atoms are added at the final stage);
- L-time: the average time (s) to construct an ANN over all $10 \times 5 = 50$ ANNs;
- t-$\mathrm{R}^2_{\mathrm{cv}}$ (best): the best value of t-$\mathrm{R}^2_{\mathrm{cv}}(j)$ over all architectures A_j, $j \in [1, 10]$;
- t-$\mathrm{R}^2_{\mathrm{max}}$: the maximum of $\mathrm{R}^2(f_{\mathrm{fc}}, \eta_{\mathcal{N}(j,i)}, D_\pi^{(i)})$ over all $j \in [1, 10], i \in [1, 5]$; and
- Arch.: The architecture A_j, $j \in [1, 10]$ that attains t-$\mathrm{R}^2_{\mathrm{max}}$. An architecture $(K, p, 1)$ (resp., $(K, p_1, p_2, 1)$) consists of an input layer with K nodes, a hidden layer with p nodes (resp., two hidden layers with p_1 and p_2 nodes, respectively), and an output layer with a single node, where K is equal to the number of descriptors in the feature vector.

From Table 4, we see that the execution of Stage 3 was considerably successful, where most of t-$\mathrm{R}^2_{\mathrm{max}}$ are around 0.85 to 0.95 for all six chemical properties.

Table 4. Results of Stages 2 and 3 in Phase 1.

π	Λ	L-Time	t-R^2_{cv} (Best)	t-R^2_{max}	Arch.
Kow	C,O,N	0.7	0.959	0.983	(156,10,10,1)
Kow	C,O,N,S,Cl	0.7	0.947	0.968	(199,20,10,1)
Bp	C,O,N	3.5	0.858	0.923	(135,30,20,1)
Bp	C,O,N,S,Cl	3.3	0.821	0.899	(163,10,1)
Mp	C,O,N	3.8	0.784	0.893	(139,40,1)
Mp	C,O,N,S,Cl	4.1	0.796	0.880	(170,10,10,1)
Fp	C,O,N	1.1	0.750	0.874	(128,40,1)
Fp	C,O,N,S,Cl	1.8	0.707	0.853	(157,10,10,1)
Lp	C,O,N	0.5	0.868	0.908	(121,30,1)
Lp	C,O,N,S,Cl	0.7	0.861	0.892	(137,20,10,1)
Sl	C,O,N	0.7	0.870	0.913	(159,30,1)
Sl	C,O,N,S,Cl	0.9	0.870	0.903	(201,30,20,1)

An Additional Experiment in Stage 3. We conducted an additional experiment to compare our new feature function f_{fc} with the feature function f_{ec} based edge-configuration in the previous method [27] designed with the same framework. Note that the previous feature vector $f_{ec}(G)$ can be defined only for a cyclic graph G, whereas our feature vector $f_{fc}(G)$ is defined for an arbitrary graph G. For each property $\pi \in \{$Kow, Bp, Mp, Fp, Lp, Sl$\}$, we set a set Λ of chemical elements to be $\{\mathtt{C}, \mathtt{O}, \mathtt{N}, \mathtt{S}, \mathtt{Cl}\}$ and then collect a data set $\widetilde{D}_\pi$ of chemical cyclic graphs from the data set D_π of all chemical graphs over the set Λ of chemical elements in the previous experiment. For each of the feature functions f_{ec} and f_{fc}, we constructed five prediction functions with the same set of ten architectures A_j, $j \in [1, 10]$ and the data set $\widetilde{D}_\pi$ of chemical cyclic graphs in the same manner of the previous experiment.

Table 5 shows the results of this experiment, where the table also includes the result of prediction functions by f_{fc} in the set D_π of all chemical graphs. In the table, we denote the following:

- $|\widetilde{D}_\pi|$, $|D_\pi|$: the size of data set $\widetilde{D}_\pi$ of cyclic graphs (resp., D_π of all chemical graphs) for property π;
- t-R^2_{cv} (ave.): the average of $R^2(f, \eta_{\mathcal{N}(j,i)}, D^{(i)})$ over all $j \in [1, 10], i \in [1, 5]$ for $f = f_{ec}, f_{fc}$ and $D = \widetilde{D}_\pi, D_\pi$; and
- t-R^2_{cv} (best): $\max_{j\in[1,10]}\{$the average of $R^2(f_{fc}, \eta_{\mathcal{N}(j,i)}, D^{(i)}_\pi)$ over all $i \in [1, 5]\}$.

From Table 5, we see that the score of R^2 of the prediction function by f_{fc} in chemical cyclic graphs (resp., in all chemical graphs) is improved from that by f_{ec} for properties Mp and Fp (resp., Bp, Mp, and Fp). Recall that our new feature function f_{fc} can be defined for arbitrary graphs and we can select a larger data set than that by f_{ec} in a learning stage. This advantage is observed in the experiment. We guess that the better prediction function for Bp (resp., Fp) is obtained by using f_{fc} because the size of data set becomes considerably larger from $|\widetilde{D}_\pi| = 224$ to $|D_\pi| = 425$ (resp., from $|\widetilde{D}_\pi| = 218$ to $|D_\pi| = 399$).

Table 5. Results of prediction functions by f_{ec} and f_{fc} in data set $\widetilde{D}_\pi$ of cyclic graphs and f_{fc} in data set D_π of all graphs.

		$f = f_{ec}, D = \widetilde{D}_\pi$		$f = f_{fc}, D = \widetilde{D}_\pi$			$f = f_{fc}, D = D_\pi$	
π	$\lvert\widetilde{D}_\pi\rvert$	t-R^2_{cv} (ave.)	t-R^2_{cv} (Best)	t-R^2_{cv} (ave.)	t-R^2_{cv} (Best)	$\lvert D_\pi\rvert$	t-R^2_{cv} (ave.)	t-R^2_{cv} (Best)
Kow	580	0.952	0.959	0.950	0.954	837	0.944	0.947
Bp	224	0.688	0.718	0.680	0.693	425	0.809	0.821
Mp	348	0.668	0.694	0.712	0.736	548	0.776	0.796
Fp	218	0.435	0.476	0.574	0.623	399	0.688	0.707
Lp	776	0.832	0.842	0.853	0.861	779	0.854	0.861
Sl	638	0.851	0.863	0.853	0.861	847	0.860	0.870

Results on Phase 2.

We prepared the following instances (a–d) for conducting experiments of Stages 4 and 5 in Phase 2.

(a) $I_{\rm a} = (G_{\rm C}, \sigma_{\rm int}, \sigma_{\rm ce})$: The instance used in Section 2.2 to explain the target specification.

(b) $I_{{\rm b},i} = (G^i_{\rm C}, \sigma^i_{\rm int}, \sigma^i_{\rm ce})$, $i = 1,2,3,4$: An instance for inferring chemical graphs with rank at most 2. In the four instances $I_{{\rm b},i}$, $i = 1,2,3,4$, the following specifications in $(\sigma_{\rm int}, \sigma_{\rm ce})$ are common.

Set $\Lambda := \{\mathtt{C},\mathtt{N},\mathtt{O}\}$, set $\Lambda^{\rm int}_{\rm dg}$ to be the set of all possible symbols in $\Lambda \times [1,4]$, and set $\Gamma^{\rm int}$ to be the set of all possible edge-configurations. Set $\Lambda^*(v) := \Lambda$, $v \in V_{\rm C}$.
The lower bounds $\ell_{\rm LB}$, ${\rm bl}_{\rm LB}$, ${\rm ch}_{\rm LB}$, ${\rm bd}_{2,{\rm LB}}$, ${\rm bd}_{3,{\rm LB}}$, ${\rm na}_{\rm LB}$, ${\rm na}^{\rm int}_{\rm LB}$, ${\rm ns}^{\rm int}_{\rm LB}$, ${\rm ac}^{\rm int}_{\rm LB}$, ${\rm ec}^{\rm int}_{\rm LB}$ are all set to be 0.
The upper bounds $\ell_{\rm UB}$, ${\rm bl}_{\rm UB}$, ${\rm ch}_{\rm UB}$, ${\rm bd}_{2,{\rm UB}}$, ${\rm bd}_{3,{\rm UB}}$, ${\rm na}_{\rm UB}$, ${\rm na}^{\rm int}_{\rm UB}$, ${\rm ns}^{\rm int}_{\rm UB}$, ${\rm ac}^{\rm int}_{\rm UB}$, ${\rm ec}^{\rm int}_{\rm UB}$ are all set to be an upper bound n^* on $n(G^*)$.
For each property π, let $\mathcal{F}(D_\pi)$ denote the set of 2-fringe-trees in the compounds in D_π, and select a subset $\mathcal{F}^i_\pi \subseteq \mathcal{F}(D_\pi)$ with $|\mathcal{F}^i_\pi| = 45 - 5i$, $i \in [1,5]$. For each instance $I_{{\rm b},i}$, set $\mathcal{F}_E := \mathcal{F}(v) := \mathcal{F}^i_\pi$, $v \in V_{\rm C}$.

Instance $I_{{\rm b},1}$ is given by the rank-1 seed graph $G^1_{\rm C}$ in Figure 9a and Instances $I_{{\rm b},i}$, $i = 2,3,4$ are given by the rank-2 seed graph $G^i_{\rm C}$, $i = 2,3,4$ in Figure 9b–d.

(i) For instance $I_{{\rm b},1}$, select as a seed graph the monocyclic graph $G^1_{\rm C} = (V_{\rm C}, E_{\rm C} = E_{(\geq 2)} \cup E_{(\geq 1)})$ in Figure 9a , where $V_{\rm C} = \{u_1, u_2\}$, $E_{(\geq 2)} = \{a_1\}$ and $E_{(\geq 1)} = \{a_2\}$. Set ${\rm n}^{\rm int}_{\rm LB} := 0, {\rm n}^{\rm int}_{\rm UB} := 12$ and $n_{\rm LB} := n^* := 38$. We include a linear constraint $\ell(a_1) \leq \ell(a_2)$ as part of the side constraint.

(ii) For instance $I_{{\rm b},2}$, select as a seed graph the graph $G^2_{\rm C} = (V_{\rm C}, E_{\rm C} = E_{(\geq 2)} \cup E_{(\geq 1)} \cup E_{(=1)})$ in Figure 9b, where $V_{\rm C} = \{u_1, u_2, u_3, u_4\}$, $E_{(\geq 2)} = \{a_1, a_2\}$, $E_{(\geq 1)} = \{a_3\}$ and $E_{(=1)} = \{a_4, a_5\}$. Set ${\rm n}^{\rm int}_{\rm LB} := {\rm n}^{\rm int}_{\rm UB} := 30$ and $n_{\rm LB} := n^* := 50$. We include a linear constraint $\ell(a_1) \leq \ell(a_2)$.

(iii) For instance $I_{{\rm b},3}$, select as a seed graph the graph $G^3_{\rm C} = (V_{\rm C}, E_{\rm C} = E_{(\geq 2)} \cup E_{(\geq 1)} \cup E_{(=1)})$ in Figure 9c, where $V_{\rm C} = \{u_1, u_2, u_3, u_4\}$, $E_{(\geq 2)} = \{a_1\}$, $E_{(\geq 1)} = \{a_2, a_3\}$ and $E_{(=1)} = \{a_4, a_5\}$. Set ${\rm n}^{\rm int}_{\rm LB} := {\rm n}^{\rm int}_{\rm UB} := 30$ and $n_{\rm LB} := n^* := 50$. We include linear constraints $\ell(a_1) \leq \ell(a_2) + \ell(a_3)$ and $\ell(a_2) \leq \ell(a_3)$.

(iv) For instance $I_{{\rm b},4}$, select as a seed graph the graph $G^4_{\rm C} = (V_{\rm C}, E_{\rm C} = E_{(\geq 2)} \cup E_{(\geq 1)} \cup E_{(=1)})$ in Figure 9d, where $V_{\rm C} = \{u_1, u_2, u_3, u_4\}$, $E_{(\geq 1)} = \{a_1, a_2, a_3\}$ and $E_{(=1)} = \{a_4, a_5\}$. Set ${\rm n}^{\rm int}_{\rm LB} := {\rm n}^{\rm int}_{\rm UB} := 30$ and $n_{\rm LB} := n^* := 50$. We include linear constraints $\ell(a_2) \leq \ell(a_1) + 1$, $\ell(a_2) \leq \ell(a_3) + 1$ and $\ell(a_1) \leq \ell(a_3)$.

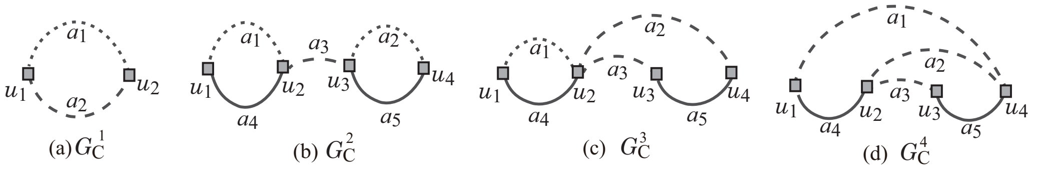


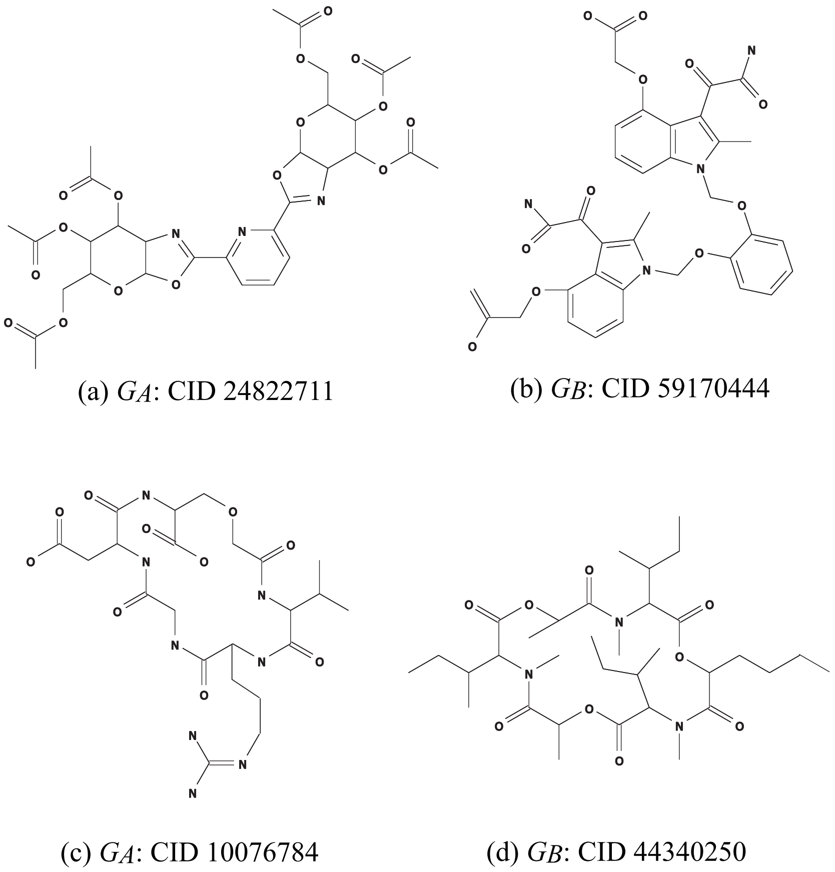
Figure 9. An illustration of seed graphs: (**a**) A monocyclic graph $G^1_{\rm C}$; (**b**) A rank-2 cyclic graph $G^2_{\rm C}$ with two vertex-disjoint cycles; (**c**) A rank-2 cyclic graph $G^3_{\rm C}$ with two disjoint cycles sharing a vertex; (**d**) A rank-2 cyclic graph $G^4_{\rm C}$ with three cycles.

We define instances in (c) and (d) in order to find chemical graphs that have an intermediate structure of given two chemical cyclic graphs $G_A = (H_A = (V_A, E_A), \alpha_A, \beta_A)$ and $G_B = (H_B = (V_B, E_B), \alpha_B, \beta_B)$. Let $\Lambda^{\rm int}_A$ and $\Lambda^{\rm int}_{{\rm dg},A}$ denote the sets of chemical elements and chemical symbols of the interior-vertices in G_A, $\Gamma^{\rm int}_A$ denote the sets of edge-configurations of the interior-edges in G_A, and $\mathcal{F}_A$ denote the set of 2-fringe-trees in G_A. Analogously define sets $\Lambda^{\rm int}_B$, $\Lambda^{\rm int}_{{\rm dg},B}$, $\Gamma^{\rm int}_B$, and $\mathcal{F}_B$ in G_B.

(c) $I_c = (G_C, \sigma_{int}, \sigma_{ce})$: An instance aimed to infer a chemical graph $G^\dagger$ such that the core of $G^\dagger$ is equal to the core of G_A and the frequency of each edge-configuration in the non-core of $G^\dagger$ is equal to that of G_B. We use chemical compounds CID 24822711 and CID 59170444 in Figure 10a,b for G_A and G_B, respectively.
Set a seed graph $G_C = (V_C, E_C = E_{(=1)})$ to be the core of G_A.
Set $\Lambda := \{\mathtt{C}, \mathtt{N}, \mathtt{O}\}$, and set Λ^{int}_{dg} to be the set of all possible chemical symbols in $\Lambda \times [1, 4]$.
Set $\Gamma^{int} := \Gamma^{int}_A \cup \Gamma^{int}_B$ and $\Lambda^*(v) := \{\alpha_A(v)\}, v \in V_C$.
Set $\mathrm{n}^{int}_{LB} := \min\{\mathrm{n}^{int}(G_A), \mathrm{n}^{int}(G_B)\}, \mathrm{n}^{int}_{UB} := \max\{\mathrm{n}^{int}(G_A), \mathrm{n}^{int}(G_B)\}$,
$n_{LB} := \min\{n(G_A), n(G_B)\} - 10$ and $n^* := \max\{n(G_A), n(G_B)\} + 5$.
Set lower bounds ℓ_{LB}, bl_{LB}, ch_{LB}, $\mathrm{bd}_{2,LB}$, $\mathrm{bd}_{3,LB}$, na_{LB}, na^{int}_{LB}, ns^{int}_{LB} and ac^{int}_{LB} to be 0.
Set upper bounds ℓ_{UB}, bl_{UB}, ch_{UB}, $\mathrm{bd}_{2,UB}$, $\mathrm{bd}_{3,UB}$, na_{UB}, na^{int}_{UB}, ns^{int}_{UB} and ac^{int}_{UB} to be n^*.
Set $\mathrm{ec}^{int}_{LB}(\gamma)$ to be the number of core-edges in G_A with $\gamma \in \Gamma^{int}$ and $\mathrm{ec}^{int}_{UB}(\gamma)$ to be the number interior-edges in G_A and G_B with edge-configuration γ.
Let $\mathcal{F}^{(p)}_B, p \in [1, 2]$ denote the set of chemical rooted trees r-isomorphic p-fringe-trees in G_B.
Set $\mathcal{F}_E := \mathcal{F}(v) := \mathcal{F}^{(1)}_B \cup \mathcal{F}^{(2)}_B, v \in V_C$.

(d) $I_d = (G^1_C, \sigma_{int}, \sigma_{ce})$: An instance aimed to infer a chemical monocyclic graph $G^\dagger$ such that the frequency vector of edge-configurations in $G^\dagger$ is a vector obtained by merging those of G_A and G_B. We use chemical monocyclic compounds CID 10076784 and CID 44340250 in Figure 10c,d for G_A and G_B, respectively. Set a seed graph to be the monocyclic seed graph $G^1_C = (V_C, E_C = E_{(\geq 2)} \cup E_{(\geq 1)})$ with $V_C = \{u_1, u_2\}$, $E_{(\geq 2)} = \{a_1\}$ and $E_{(\geq 1)} = \{a_2\}$ in Figure 9a.
Set $\Lambda := \{\mathtt{C}, \mathtt{N}, \mathtt{O}\}$, $\Lambda^{int}_{dg} := \Lambda^{int}_{dg,A} \cup \Lambda^{int}_{dg,B}$ and $\Gamma^{int} := \Gamma^{int}_A \cup \Gamma^{int}_B$.
Set $\mathrm{n}^{int}_{LB} := \min\{\mathrm{n}^{int}(G_A), \mathrm{n}^{int}(G_B)\}, \mathrm{n}^{int}_{UB} := \max\{\mathrm{n}^{int}(G_A), \mathrm{n}^{int}(G_B)\}$,
$n_{LB} := \min\{n(G_A), n(G_B)\}$ and $n^* := \max\{n(G_A), n(G_B)\}$.
Set lower bounds ℓ_{LB}, bl_{LB}, ch_{LB}, $\mathrm{bd}_{2,LB}$, $\mathrm{bd}_{3,LB}$, na_{LB}, na^{int}_{LB}, ns^{int}_{LB} and ac^{int}_{LB} to be 0.
Set upper bounds ℓ_{UB}, bl_{UB}, ch_{UB}, $\mathrm{bd}_{2,UB}$, $\mathrm{bd}_{3,UB}$, na_{UB}, na^{int}_{UB}, ns^{int}_{UB} and ac^{int}_{UB} to be n^*.
For each edge-configuration $\gamma \in \Gamma^{int}$, let $\boldsymbol{x}^*_A(\gamma^{int})$ (resp., $\boldsymbol{x}^*_B(\gamma^{int})$) denote the number of interior-edges with γ in G_A (resp., G_B), $\gamma \in \Gamma^{int}$ and set
$\boldsymbol{x}^*_{min}(\gamma) := \min\{\boldsymbol{x}^*_A(\gamma), \boldsymbol{x}^*_B(\gamma)\}, \boldsymbol{x}^*_{max}(\gamma) := \max\{\boldsymbol{x}^*_A(\gamma), \boldsymbol{x}^*_B(\gamma)\}$,
$\mathrm{ec}^{int}_{LB}(\gamma) := \lfloor (3/4)\boldsymbol{x}^*_{min}(\gamma) + (1/4)\boldsymbol{x}^*_{max}(\gamma) \rfloor$ and
$\mathrm{ec}^{int}_{UB}(\gamma) := \lceil (1/4)\boldsymbol{x}^*_{min}(\gamma) + (3/4)\boldsymbol{x}^*_{max}(\gamma) \rceil$.
Set $\mathcal{F}_E := \mathcal{F}(v) := \mathcal{F}_A \cup \mathcal{F}_B, v \in V_C$.

In Stage 5, before we formulate an MILP for inferring a target chemical graph $G^\dagger$ for each instance I, we reduce the input layer of an ANN $\mathcal{N}$ constructed in Stage 3 so that the input layer consists of input nodes that correspond to the descriptors actually used in the specification $(G_C, \sigma_{int}, \sigma_{ce})$ of the instance I, i.e., we remove any input nodes in $\mathcal{N}$ that represent the frequency of edge-configurations in $\Gamma^{int}(D_\pi)$ and chemical rooted trees $\psi \in \mathcal{F}(D_\pi)$ not contained in the specification $(G_C, \sigma_{int}, \sigma_{ce})$ of I. For example, there are $|\mathcal{F}(D_\pi)| = 109$ chemical rooted trees in the set of 2-fringe-trees in the data set D_π with π = Kow in Table 3, and an ANN $\mathcal{N}$ constructed in Stage 3 contains 109 input nodes that correspond to the descriptors for the fringe-configuration. However, the set of input nodes for the fringe-configuration is reduced to a set of $|\mathcal{F}^*| = 40$ input nodes when we formulate an MILP for solving instance $I_{b,1}$, saving the number of integer variables.



(a) G_A: CID 24822711 (b) G_B: CID 59170444

(c) G_A: CID 10076784 (d) G_B: CID 44340250

Figure 10. An illustration of chemical compounds for instances I_c and I_d: (**a**) G_A: CID 24822711; (**b**) G_B: CID 59170444; (**c**) G_A: CID 10076784; (**d**) G_B: CID 44340250.

Table 6 shows the features of the seven test instances, where we denote the following:

- Λ: the set of non-hydrogen chemical elements for inferring a target graph;
- $|\Gamma^{\text{int}}|$: the number of different edge-configurations of interior-edges for inferring a target graph;
- $|\mathcal{F}^*|$: the number of different chemical rooted trees in the set $\mathcal{F}^* = \mathcal{F}_E \cup \bigcup_{v \in V_C} \mathcal{F}(v)$; and
- $[\mathrm{n}^{\text{int}}_{\text{LB}}, \mathrm{n}^{\text{int}}_{\text{UB}}]$, $[n_{\text{LB}}, n^*]$: the lower and upper bounds on $\mathrm{n}^{\text{int}}(G^\dagger)$ and $n(G^\dagger)$ for inferring a target graph $G^\dagger$.

Table 6. Features of test instances.

Instance	Λ	$\lvert\Gamma^{\text{int}}\rvert$	$\lvert\mathcal{F}^*\rvert$	$[\mathbf{n}^{\text{int}}_{\text{LB}}, \mathbf{n}^{\text{int}}_{\text{UB}}]$	$[n_{\text{LB}}, n^*]$
I_{a}	C,O,N	10	11	[30,50]	[20,28]
$I_{\text{b},1}$	C,O,N	28	40	[38,38]	[6,6]
$I_{\text{b},2}$	C,O,N	28	35	[50,50]	[30,30]
$I_{\text{b},3}$	C,O,N	28	30	[50,50]	[30,30]
$I_{\text{b},4}$	C,O,N	28	25	[50,50]	[30,30]
I_{c}	C,O,N	8	12	[46,46]	[24,24]
I_{d}	C,O,N	7	8	[40,45]	[18,18]

Stage 4. To solve an MILP in Stage 4, we used `CPLEX` version 12.10. Tables 7–12 show the results on Stages 4 and 5, where we denote the following:

- $[\underline{a}, \overline{a}]$: the minimum and maximum values of $a(G)$ in π over compounds G in D_π in Table 3;
- $[\underline{y}, \overline{y}]$: $\underline{y}$ (resp., $\overline{y}$) denotes the minimum (resp., maximum) target value y with $\underline{a} \leq y \leq \overline{a}$ such that the MILP instance for the target value $y^* = y$ becomes feasible (i.e., admits

a target chemical graph $G^{\dagger}$). To determine the minimum and minimum target values $\underline{y}$ and $\overline{y}$, we solved many numbers of MILP instances. Note that the MILP instance may become infeasible for some value y within the range $[\underline{y}, \overline{y}]$;

- y^{*}: a target value in $[\underline{y}, \overline{y}]$ for a property π;
- #v: the number of variables in the MILP in Stage 4;
- #c: the number of constraints in the MILP in Stage 4;
- IP-time: the time (sec.) to solve the MILP in Stage 4;
- n: the number $n(G^{\dagger})$ of non-hydrogen atoms in the chemical graph $G^{\dagger}$ inferred in Stage 4; and
- $\mathrm{n}^{\mathrm{int}}$: the number $\mathrm{n}^{\mathrm{int}}(G^{\dagger})$ of interior-vertices in the chemical graph $G^{\dagger}$ inferred in Stage 4.

Figure 11a illustrates the chemical graph $G^{\dagger}$ inferred from instance I_{c} with $y^{*} = 3.0$ of Kow in Table 7.

Table 7. Results of Stage 4 for Kow.

Instance	$[\underline{a}, \overline{a}]$	$[\underline{y}, \overline{y}]$	y^{*}	#v	#c	IP-Time	n	$\mathrm{n}^{\mathrm{int}}$
I_{a}	[−7.53, 13.45]	[−7.0, 13.4]	3.2	7663	9162	3.9	35	24
$I_{\mathrm{b,1}}$	[−7.53, 13.45]	[−7.5, 13.4]	3.0	9894	6626	17.5	38	7
$I_{\mathrm{b,2}}$	[−7.53, 13.45]	[−7.5, 13.4]	3.0	11,514	8934	14.0	50	30
$I_{\mathrm{b,3}}$	[−7.53, 13.45]	[−7.5, 13.4]	3.0	11,318	8926	24.6	50	30
$I_{\mathrm{b,4}}$	[−7.53, 13.45]	[−7.5, 13.4]	3.0	11,122	8918	22.0	50	30
I_{c}	[−7.53, 13.45]	[−7.5, 13.4]	3.0	7867	8630	2.1	49	32
I_{d}	[−7.53, 13.45]	[−7.5, 13.4]	3.0	5395	6899	5.2	45	23

Table 8. Results of Stage 4 for Bp.

Instance	$[\underline{a}, \overline{a}]$	$[\underline{y}, \overline{y}]$	y^{*}	#v	#c	IP-Time	n	$\mathrm{n}^{\mathrm{int}}$
I_{a}	[−11.70, 470.0]	[352, 470]	411	7583	8982	2.7	42	25
$I_{\mathrm{b,1}}$	[−11.70, 470.0]	[−11, 470]	229	9816	6449	2.7	38	7
$I_{\mathrm{b,2}}$	[−11.70, 470.0]	[−11, 470]	229	11,436	8757	9.1	50	30
$I_{\mathrm{b,3}}$	[−11.70, 470.0]	[−11, 470]	229	11,240	8749	11.0	50	30
$I_{\mathrm{b,4}}$	[−11.70, 470.0]	[−11, 470]	229	11,044	8741	24.0	50	30
I_{c}	[−11.70, 470.0]	[170, 470]	320	7575	8450	25.9	49	33
I_{d}	[−11.70, 470.0]	[151, 470]	310	5315	6719	4.4	43	23

Table 9. Results of Stage 4 for Mp.

Instance	$[\underline{a}, \overline{a}]$	$[\underline{y}, \overline{y}]$	y^{*}	#v	#c	IP-Time	n	$\mathrm{n}^{\mathrm{int}}$
I_{a}	[−185.3, 300.0]	[55, 300]	177.5	7602	9023	16.1	41	24
$I_{\mathrm{b,1}}$	[−185.3, 300.0]	[−180, 300]	60	9833	6487	2.3	38	9
$I_{\mathrm{b,2}}$	[−185.3, 300.0]	[−185, 300]	57.4	11,453	8795	44.7	50	30
$I_{\mathrm{b,3}}$	[−185.3, 300.0]	[−185, 300]	57.4	11,257	8787	10.5	50	30
$I_{\mathrm{b,4}}$	[−185.3, 300.0]	[−185, 300]	57.4	11,061	8779	93.9	50	30
I_{c}	[−185.3, 300.0]	[253, 300]	260.0	7580	6172	24.0	41	33
I_{d}	[−185.3, 300.0]	[−75, 299]	58	5110	4050	104.6	45	23

Table 10. Results of Stage 4 for FP.

Instance	$[\underline{a}, \overline{a}]$	$[\underline{y}, \overline{y}]$	y^*	#V	#C	IP-Time	n	$\mathrm{n}^{\mathrm{int}}$
I_{a}	[−82.99, 300.0]	[98, 300]	199	7459	8696	1.6	35	22
$I_{\mathrm{b},1}$	[−82.99, 300.0]	[−82, 300]	109	9694	6166	1.4	38	8
$I_{\mathrm{b},2}$	[−82.99, 300.0]	[−82, 300]	109	11,314	8474	8.7	50	30
$I_{\mathrm{b},3}$	[−82.99, 300.0]	[−82, 300]	109	11,118	8466	25.8	50	30
$I_{\mathrm{b},4}$	[−82.99, 300.0]	[−82, 300]	109	10,922	8458	8.5	50	30
I_{c}	[−82.99, 300.0]	[250, 300]	275	7667	8170	60.9	47	34
I_{d}	[−82.99, 300.0]	[54, 300]	177	5193	6436	2.0	45	23

Table 11. Results of Stage 4 for LP.

Instance	$[\underline{a}, \overline{a}]$	$[\underline{y}, \overline{y}]$	y^*	#V	#C	IP-Time	n	$\mathrm{n}^{\mathrm{int}}$
I_{a}	[−3.6, 6.84]	[−3.6, 6.8]	1.6	7597	9008	1.9	39	23
$I_{\mathrm{b},1}$	[−3.6, 6.84]	[−3.6, 6.8]	1.6	9836	6481	2.9	38	8
$I_{\mathrm{b},2}$	[−3.6, 6.84]	[−3.6, 6.8]	1.6	11,456	8789	21.1	50	30
$I_{\mathrm{b},3}$	[−3.6, 6.84]	[−3.6, 6.8]	1.6	11,260	8781	20.4	50	30
$I_{\mathrm{b},4}$	[−3.6, 6.84]	[−3.6, 6.8]	1.6	11,064	8773	24.2	50	30
I_{c}	[−3.6, 6.84]	[−3.6, 6.8]	1.6	7801	8476	1.1	47	32
I_{d}	[−3.6, 6.84]	[−3.6, 6.8]	1.6	5335	6754	4.3	45	23

Figure 11b illustrates the chemical graph $G^{\dagger}$ inferred from instance I_{d} with $y^* = 1.6$ of LP in Table 11.

Table 12. Results of Stage 4 for SL.

Instance	$[\underline{a}, \overline{a}]$	$[\underline{y}, \overline{y}]$	y^*	#V	#C	IP-Time	n	$\mathrm{n}^{\mathrm{int}}$
I_{a}	[−9.33, 1.11]	[−9.3, −2.0]	−5.6	7674	9186	2.4	41	23
$I_{\mathrm{b},1}$	[−9.33, 1.11]	[−9.3, −2.0]	−5.6	9906	6650	22.3	38	12
$I_{\mathrm{b},2}$	[−9.33, 1.11]	[−9.3, −2.0]	−5.6	11,526	8958	15.2	50	30
$I_{\mathrm{b},3}$	[−9.33, 1.11]	[−9.3, −2.0]	−5.6	11,330	8950	16.2	50	30
$I_{\mathrm{b},4}$	[−9.33, 1.11]	[−9.3, −2.0]	−5.6	11,134	8942	122.7	50	30
I_{c}	[−9.33, 1.11]	[−9.3, −2.0]	−5.6	7874	8648	1.2	54	33
I_{d}	[−9.33, 1.11]	[−9.3, −3.0]	−6.1	5402	6917	8.1	43	23

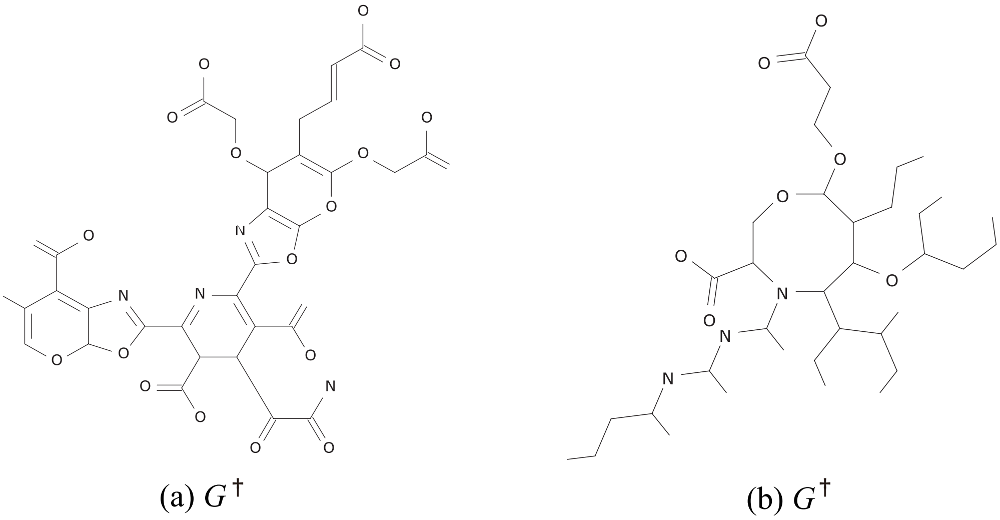

(a) $G^{\dagger}$ (b) $G^{\dagger}$

Figure 11. (**a**) $G^{\dagger}$ inferred from I_{c} with $y^* = 3.0$ of KOW; (**b**) $G^{\dagger}$ inferred from I_{d} with $y^* = 1.6$ of LP.

The topological specification of instances I_a, I_c and I_d is more restricted than that of the other instances, and thereby the feasible target range $[\underline{y}, \overline{y}]$ of I_a, I_c and I_d is rather narrower than the original range $[\underline{a}, \overline{a}]$ for some property π. We see that the running time for solving an MILP instance with $n = 50$ is 8.5 to 122 (s), which is much smaller than the running time of 61 to 12058 (s) to solve a similar set of MILP instances with $n = 50$ in the experimental result for the previous method [28].

Stage 5. We computed chemical isomers G^* of each target chemical graph $G^\dagger$ inferred in Stage 4. We execute the algorithm for generating chemical isomers of $G^\dagger$ up to 100 when the number of all chemical isomers exceeds 100. The algorithm can evaluate a lower bound on the total number of all chemical isomers $G^\dagger$ without generating all of them.

Tables 13 and 14 show the computational results of the experiment, where we denote the following:

- DP-time: the running time (s) to execute the dynamic programming algorithm in Stage 5 to compute a lower bound on the number of all chemical isomers G^* of $G^\dagger$ and generate all (or up to 100) chemical isomers G^*;
- G-LB: a lower bound on the number of all chemical isomers G^* of $G^\dagger$; and
- #G: the number of all (or up to 100) chemical isomers G^* of $G^\dagger$ generated in Stage 5.

From Tables 13 and 14, we observe that the running time for generating up to 100 target chemical graphs in Stage 5 is not considerably larger than that in Stage 4.

Table 13. Results of Stage 5 for Kow, Lp, and Bp.

		Kow			Lp			Bp	
Instance	**DP-Time**	**G-LB**	**#G**	**DP-Time**	**G-LB**	**#G**	**DP-time**	**G-LB**	**#G**
I_a	0.031	16	16	0.164	128	100	0.164	1.4×10^5	100
I_b^1	0.149	2.8×10^5	100	0.148	2.0×10^{10}	100	0.162	4.4×10^5	100
I_b^2	44.1	3.9×10^{10}	100	118	900	100	171	6	6
I_b^3	27.2	20	20	80.2	6	6	28.6	7	7
I_b^4	0.166	6000	100	73	12	12	142	5	5
I_c	0.166	6000	100	0.168	288	100	0.168	4.0×10^5	100
I_d	22.3	8.3×10^{10}	100	1.44	3.2×10^8	100	1.7	9.7×10^9	100

Table 14. Results of Stage 5 for Fp, Mp, and Sl.

		Fp			Mp			Sl	
Instance	**DP-Time**	**G-LB**	**#G**	**DP-Time**	**G-LB**	**#G**	**DP-Time**	**G-LB**	**#G**
I_a	0.057	32	32	0.165	256	100	0.165	1024	100
I_b^1	0.164	3.1×10^6	100	0.166	1.4×10^6	100	0.163	4.5×10^5	100
I_b^2	28.8	720	100	8.26	2.4×10^{10}	100	1.07	5.6×10^9	100
I_b^3	72.2	27	27	51.9	1	1	46.5	1680	100
I_b^4	40.3	20	20	125	6.1×10^7	100	7.01	1.1×10^8	100
I_c	0.169	1.1×10^5	100	0.173	6048	100	0.168	120	100
I_d	0.057	32	32	0.17	4.2×10^8	100	0.165	1024	100

4. Discussions and Conclusions

The framework of designing chemical graphs using ANNs and MILP has been proposed [23] as a basis of a total system of the QSAR and the inverse of QSAR, where the inverse of a prediction function produced by an ANN is solved by an MILP. The merit of the framework is that the inverse problem can be treated exactly as a mathematical problem, and an MILP instance with a moderate size can be efficiently solved with a fast MILP solver. On the other hand, the main technical concern in applying the framework is in defining a feature vector of a chemical graph in terms of graph theoretical descriptors so that the computation of a feature vector can be simulated with a set of linear constraints in an MILP. So far, the framework has been applied to the design of new methods of inferring several restricted classes of chemical graphs such as the graphs with rank at most 2 and the ρ-lean cyclic graphs [26,28].

Herein, we examine some technical issues in the previous method before we observe some new features of our method in this paper.

In the feature vector of the previous models [26,28], the structure of subgraphs used as descriptors is only a pair of adjacent vertices, called adjacency-configuration or edge-configuration, which is significantly limited from a variety of subgraphs used in a more sophisticated construction of a feature vector such as the fingerprint. However, including the occurrence of a certain subgraph with only a few vertices as part of a feature vector may require realizing a mechanism of the subgraph isomorphism in an MILP that simulates the computation of such an occurrence and can easily make the resulting MILP very complicated and hard to solve. Furthermore, the feature vector can be defined only for cyclic graphs and we need to eliminate any acyclic graphs from the original data set before we construct a prediction function. This may reduce a data set to an unnecessarily small size or reduce the chances of capturing important information on QSAR over all types of graphs.

A branch-parameter ρ was originally introduced as a new measure to the "agglomeration degree" of trees [24] and then used to define restricted classes of acyclic and cyclic graphs [24,27]. In fact, such a restriction on the structure of target chemical graphs was rather necessary to reduce the size of an MILP formulation that simulates a selection process of a target chemical graph from a supergraph (called a scheme graph), where the number of variables and constraints required to infer a chemical graph with n^* non-hydrogen atoms is $O(n^*)$ when some other parameters such as ρ are regarded as constants.

Although nearly 97% of cyclic chemical compounds with up to 100 non-hydrogen atoms in PubChem are 2-lean [24], the way of specifying the topological structure of a target chemical graph in the previous method [26,28] was based on the core and the non-core of a chemical graph, and we could not include a fixed substructure in the non-core of a target chemical graph.

Compared with the previous models, the two-layered model proposed in this paper is rather simple, where a chemical graph is regarded as a combination of the interior and the exterior. The new model can deal with chemical compounds with any graph structure and include a prescribed structure in both of the acyclic and cyclic parts of a target chemical graph as long as the requirement on target chemical graphs is described under the set of specification rules introduced in this paper. This considerably improves the availability of the framework in a practical application.

The feature vector of our two-layered model can be defined for arbitrary graphs. In the new feature vector, the exterior of a chemical graph is encoded into fringe-configurations, i.e., the occurrence of each chemical rooted tree with height at most ρ, where we may regard that the set of such a chemical rooted trees plays a similar role of some types of functional groups. In our method, we include as part of the descriptors of a feature vector the occurrence of each of such chemical rooted trees and the descriptors of our feature vector on the exterior of a chemical graph may have an analogous effect with the fingerprint.

Our specification of target chemical graphs can specify a candidate set $\mathcal{F}$ of chemical rooted trees that are allowed to be used as chemical rooted trees in the exterior of a target chemical graph. This allows us to control the chemical property of target chemical graphs in a more meaningful way since chemical properties of some rooted trees in $\mathcal{F}$ are known as functional groups and some kinds of rooted trees can be prohibited in a target chemical graph, if necessary, just by excluding from the candidate set $\mathcal{F}$. Although the number $|\mathcal{F}(D_\pi)|$ of different kinds of such chemical trees in a data set D_π from PubChem is approximately up to 300 for $\rho = 2$ in many cases and the number of input nodes in an ANN $\mathcal{N}$ becomes over $|\mathcal{F}(D_\pi)|$, we derived an MILP formulation for inferring a chemical graph with with n^* non-hydrogen atoms and a candidate set $\mathcal{F}$ of chemical rooted trees by using

$O(n^* + |\mathcal{F}|)$ variables and $O(n^*|\mathcal{F}|)$ constraints when the number of interior-vertices is constant, where $|\mathcal{F}|$ can be quite small compared with $|\mathcal{F}(D_\pi)|$.

We have implemented the proposed method for inferring chemical compounds with a prescribed topological substructure setting $\rho = 2$. The results of computational experiments using some chemical properties such as octanol/water partition coefficient, boiling point, melting point, flash point, lipophylicity, and solubility suggest that the proposed system can infer chemical graphs with 50 non-hydrogen atoms.

For a larger branch-parameter, say $\rho = 3, 4$, we obtain a more variety of chemical rooted trees which provides new descriptors in a feature vector and new candidates for fringe-trees in the exterior in a target chemical graph, whereas the number of different chemical rooted trees in $\mathcal{F}(D_\pi)$ may increase rapidly.

It is left as a future work to use other learning methods such as decision tree, random forest, graph convolution, and an ensemble method in Stages 3 and 4 in the framework.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1.

Author Contributions: Conceptualization, H.N. and T.A.; methodology, H.N.; software, N.A.A., J.Z., Y.S., K.H., and L.Z.; validation, N.A.A., J.Z., and H.N.; formal analysis, H.N.; data resources, L.Z., K.H., H.N., and T.A.; writing—original draft preparation, H.N.; writing—review and editing, N.A.A. and T.A.; project administration, H.N.; funding acquisition, T.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported, in part, by Japan Society for the Promotion of Science, Japan, under Grant #18H04113.

Data Availability Statement: Source code of the implementation of our algorithm is freely available from https://github.com/ku-dml/mol-infer.

Acknowledgments: This research was supported, in part, by Japan Society for the Promotion of Science, Japan, under Grant #18H04113.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

ANN	artificial neural network
MILP	mixed integer linear programming

References

1. Miyao, T.; Kaneko, H.; Funatsu, K. Inverse QSPR/QSAR analysis for chemical structure generation (from y to x). *J. Chem. Inf. Model.* **2016**, *56*, 286–299. [CrossRef]
2. Ikebata, H.; Hongo, K.; Isomura, T.; Maezono, R.; Yoshida, R. Bayesian molecular design with a chemical language model. *J. Comput. Aided Mol. Des.* **2017**, *31*, 379–391. [CrossRef]
3. Rupakheti, C.; Virshup, A.; Yang, W.; Beratan, D.N. Strategy to discover diverse optimal molecules in the small molecule universe. *J. Chem. Inf. Model.* **2015**, *55*, 529–537. [CrossRef] [PubMed]
4. Fujiwara, H.; Wang, J.; Zhao, L.; Nagamochi, H.; Akutsu, T. Enumerating treelike chemical graphs with given path frequency. *J. Chem. Inf. Model.* **2008**, *48*, 1345–1357. [CrossRef]
5. Kerber, A.; Laue, R.; Grüner, T.; Meringer, M. MOLGEN 4.0. *MATCH Commun. Math. Comput. Chem.* **1998**, 37, 205–208.
6. Li, J.; Nagamochi, H.; Akutsu, T. Enumerating substituted benzene isomers of tree-like chemical graphs. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2016**, *15*, 633–646. [CrossRef]
7. Reymond, J.L. The chemical space project. *Acc. Chem. Res.* **2015**, *48*, 722–730. [CrossRef] [PubMed]
8. Bohacek, R.S.; McMartin, C.; Guida, W.C. The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev.* **1996**, *16*, 3–50. [CrossRef]
9. Akutsu, T.; Fukagawa, D.; Jansson, J.; Sadakane, K. Inferring a graph from path frequency. *Discrete Appl. Math.* **2012**, *160*, 1416–1428. [CrossRef]
10. Kipf, T. N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
11. Gómez-Bombarelli, R.; Wei, J.N.; Duvenaud, D.; Hernández-Lobato, J.M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T.D.; Adams, R.P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276. [CrossRef] [PubMed]

12. Segler, M.H.S.; Kogej, T.; Tyrchan, C.; Waller, M.P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **2017**, *4*, 120–131. [CrossRef]
13. Yang, X.; Zhang, J.; Yoshizoe, K.; Terayama, K.; Tsuda, K. ChemTS: An efficient python library for de novo molecular generation. *Sci. Technol. Adv. Mater.* **2017**, *18*, 972–976. [CrossRef] [PubMed]
14. Kusner, M.J.; Paige, B.; Hernández-Lobato, J.M. Grammar variational autoencoder. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; Volume 70, pp. 1945–1954.
15. De Cao, N.; Kipf, T. MolGAN: An implicit generative model for small molecular graphs. *arXiv* **2018**, arXiv:1805.11973.
16. Madhawa, K.; Ishiguro, K.; Nakago, K.; Abe, M. GraphNVP: an invertible flow model for generating molecular graphs. *arXiv* **2019**, arXiv:1905.11600.
17. Shi, C.; Xu, M.; Zhu, Z.; Zhang, W.; Zhang, M.; Tang, J. GraphAF: A flow-based autoregressive model for molecular graph generation. *arXiv* **2020**, arXiv:2001.09382.
18. Cherkasov, A.; Muratov, E.M.N.; Fourches, D.; Varnek, A.; Baskin, I.I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y.C.; Todeschini, R.; et al. QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* **2014**, *57*, 4977–5010. [CrossRef]
19. Cramer, R.D., III; Patterson, D.E.; Bunce, J.D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967. [CrossRef]
20. Cramer, R.D. Template CoMFA generates single 3D-QSAR models that, for twelve of twelve biological targets, predict all ChEMBL-tabulated affinities. *PLoS ONE* **2015**, *10*, e0129307. [CrossRef] [PubMed]
21. Moriwaki, H.; Tian, Y-S.; Kawashita, N.; Takagi, T. Three-dimensional classification structure–activity relationship analysis using convolutional neural network. *Chem. Pharm. Bull.* **2019**, *67*, 426–432. [CrossRef]
22. Azam, N.A.; Chiewvanichakorn, R.; Zhang, F.; Shurbevski, A.; Nagamochi, H.; Akutsu, T. A method for the inverse QSAR/QSPR based on artificial neural networks and mixed integer linear programming. In Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies, Valletta, Malta, 24–26 February 2020; Volume 3, pp. 101–108.
23. Zhang, F.; Zhu, J.; Chiewvanichakorn, R.; Shurbevski, A.; Nagamochi, H.; Akutsu, T. A new integer linear programming formulation to the inverse QSAR/QSPR for acyclic chemical compounds using skeleton trees. In Proceedings of the 33rd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Kitakyushu, Japan, 22–25 September 2020; pp. 433–444.
24. Azam, N.A.; Zhu, J.; Sun, Y.; Shi, Y.; Shurbevski, A.; Zhao, L.; Nagamochi, H.; Akutsu, T. A novel method for inference of acyclic chemical compounds with bounded branch-height based on artificial neural networks and integer programming. **2020**, submitted.
25. Ito, R.; Azam, N.A.; Wang, C.; Shurbevski, A.; Nagamochi, H.; Akutsu, T. A novel method for the inverse QSAR/QSPR to monocyclic chemical compounds based on artificial neural networks and integer programming. In Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP 2020), Las Vegas, NV, USA, 27–30 July 2020.
26. Zhu, J.; Wang, C.; Shurbevski, A.; Nagamochi, H.; Akutsu, T. A novel method for inference of chemical compounds of cycle index two with desired properties based on artificial neural networks and integer programming. *Algorithms* **2020**, *13*, 124. [CrossRef]
27. Akutsu, T.; Nagamochi, H. A novel method for inference of chemical compounds with prescribed topological substructures based on integer programming. *arXiv* **2020**, arXiv: 2010.09203.
28. Zhu, J.; Azam, N.A.; Zhang, F.; Shurbevski, A.; Haraguchi, K.; Zhao, L.; Nagamochi, H.; Akutsu, T. A novel method for inferring of chemical compounds with prescribed topological substructures based on integer programming. **2020**, submitted.
29. PubChem. Available online: https://pubchem.ncbi.nlm.nih.gov/ (accessed on 13 May 2020).
30. Figshare. Available online: https://figshare.com/articles/dataset/Lipophilicity_Dataset_-_logD7_4_of_1_130_Compounds/5596750/1 (accessed on 13 May 2020).
31. A Benchmark for Molecular Machine Learning. Available online: http://moleculenet.ai/datasets-1 (accessed on 13 May 2020).