【발명의 설명】

【발명의 명칭】

제어형 그래디언트-베이즈 흐름 네트워크 기반 3차원 분자 구조 생성 시스템 및 그 방법{SYSTEM AND METHOD FOR THREE-DIMENSIONAL MOLECULAR STRUCTURE GENERATION BASED ON A CONTROLLED GRADIENT-BAYESIAN FLOW NETWORK}

【기술분야】

<0001> 본 발명은 인공지능 기반 분자 생성 및 신약 설계 기술 분야에 관한 것으로서, 상세하게는 단백질 표적 구조 정보를 이용하여 새로운 3차원 분자 구조를 설계·생성하는 구조 기반 신약 설계(Structure-Based Drug Design, SBDD) 분야에 관한 것이다. 특히, 본 발명은 연속 변수(원자 좌표)와 범주형 변수(원자 타입)를 동시에 다루기 위한 베이즈 확률 흐름 네트워크(Bayesian Flow Network)와 목표 약효 특성(결합친화도, 선택성, 합성가능성 등)에 대한 조건부 그래디언트 유도(gradient-based conditional guidance)를 통합적으로 활용하여 분자 구조의 생성, 업데이트 및 최적화를 수행하는 인공지능 기반 3차원 분자 생성 시스템 및 방법에 관한 것이다.

【발명의 배경이 되는 기술】

<0002> 신약 개발 분야에서는 단백질 표적의 3차원 구조를 활용하여 최적의 리간드를 설계하는 구조 기반 신약 설계(SBDD) 기술이 활발히 연구되고 있다. 최근에는 확산 모델(diffusion model)과 같은 생성 모델이 분자 구조 생성에 활용되고 있으나, 샘플 공간에서 직접 원자 좌표와 원자 타입을 생성하는 방식은 생성 안정성이

낮고, 결합친화도, 선택성 및 합성가능성 등 목표 속성에 대한 조건부 제어가 제한되는 문제가 있다. 또한 분자 생성 과정에서 필수적인 범주형 변수(원자 타입)는 미분이 불가능하여, 그래디언트 기반 최적화가 어려운 한계가 존재한다.

<0003> 이와 같은 문제를 해결하기 위해 베이즈 흐름 네트워크(Bayesian Flow Network)가 제안되어 매개변수 공간에서의 안정적 업데이트와 범주형 변수의 연속 표현 변환이 가능해졌으나, 여전히 실제 신약 설계에서 요구되는 속성 기반 조건부 제어 기능이 부족하다. 따라서 단백질 구조에 적합한 3차원 분자 구조를 안정적으로 생성하면서도, 목표 속성에 대한 그래디언트 기반 최적화를 동시에 수행할 수 있는 새로운 생성 모델 기술이 요구되고 있다.

【선행기술문헌】

【특허문헌】

<0004> (특허문헌 1) 한국 등록특허공보 제10-2743467호 (2024.03.21)

【발명의 내용】

【해결하고자 하는 과제】

<0005> 본 발명이 이루고자 하는 기술적 과제는, 첫째, 서로 다른 분포에서 샘플링되는 연속적인 원자 좌표 변수와 범주형 원자 타입 변수 간의 상호작용을 종래의 유도 메커니즘이 정확하게 반영하지 못함으로써 발생하는 화학적 맥락 손실 문제를 해소할 수 있는 기술을 제공하는 것이다.

<0006> 둘째, 확산 기반 생성 모델에서 범주형 변수에 그래디언트 기반 유도를 적용하기 어려워 유도 신호가 비효율적이거나 불안정해지는 문제를 해결하고, 종래의

우회 기법이 초래하는 부자연스러운 표현 및 모델 복잡성 증가 문제를 개선할 수 있는 기술을 제공하는 것이다.

<0007> 셋째, 샘플 공간에 직접 그래디언트를 주입하는 과정에서 3차원 분자의 화학적 및 구조적 유효성이 손상되어 속성 제어가 어렵고 불안정한 분자 구조가 생성되는 문제를 방지할 수 있는 안정적 생성·업데이트 방식을 제공하는 것이다.

<0008> 본 발명이 이루고자 하는 기술적 과제는 이상에서 언급한 기술적 과제로 제한되지 않으며, 언급되지 않은 또 다른 기술적 과제들은 아래의 기재로부터 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

【과제의 해결 수단】

<0009> 상기 기술적 과제를 달성하기 위하여, 제어형 그래디언트-베이즈 흐름 네트워크 기반 3차원 분자 구조 생성 시스템에 있어서, 분자를 구성하는 연속 변수 및 범주형 변수를 수신하는 입력 변수 수신부; 수신된 상기 연속 변수 및 상기 범주형 변수를 제어형 베이즈 흐름 신경망 모델부에서 이용할 수 있는 형태로 전처리하는 데이터 전처리부; 전처리된 상기 연속 변수 및 상기 범주형 변수에 기초하여 베이즈 흐름(Bayesian Flow)과 그래디언트 기반 업데이트를 수행하여, 원자 좌표 및 원자 타입을 산출함으로써 3차원 분자 구조를 생성하는 상기 제어형 베이즈 흐름 신경망 모델부; 및 상기 생성된 3차원 분자 구조의 결과를 출력하는 결과 출력부; 를 포함한다.

<0010> 여기서, 본 발명의 실시예에 따른 상기 제어형 베이즈 흐름 신경망 모델부

는, 상기 범주형 변수를 그래디언트 업데이트에 활용할 수 있도록 연속 공간으로 변환하는 베이즈 기반 범주형 변수 변환부; 상기 연속 변수 및 상기 연속 공간으로 변환된 범주형 변수에 기초하여 산출된 조건부 그래디언트를 이용하여 파라미터를 업데이트하는 조건부 그래디언트 업데이트부; 및 업데이트된 상기 파라미터에 기초하여, 단백질 구조 및 목표 속성에 조건화된 3차원 분자 구조를 생성하는 3차원 분자 구조 생성부; 를 포함할 수 있다.

<0011> 또한, 본 발명의 실시예에 따른 상기 베이즈 기반 범주형 변수 변화부는, 원자 타입 형태의 상기 범주형 변수를 아래의 식1에 의해 연속 공간으로 변환하도록 구성될 수 있다.

<0012> [식 1] $h(\boldsymbol{\theta}_{i-1}, \boldsymbol{y}, \alpha) = Softmax(e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1})$

<0013> $\boldsymbol{\theta}_{i-1}$: 이전 단계에서의 잠재 파라미터

<0014> $\boldsymbol{y}$: 범주형 변수(예: 원자 타입)를 변환하기 위한 입력 벡터

<0015> α: 해당 단계에서의 베이즈 업데이트 계수

<0016> $e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1}$: 범주형 변수 입력과 잠재 파라미터의 결합 결과

<0017> $Softmax(\cdot)$: 입력 벡터를 확률적 연속 표현으로 변환하는 함수

<0018> 또한, 본 발명의 실시예에 따른 상기 조건부 그래디언트 업데이트부는, 단백질 구조 및 목표 속성에 기초하여 상기 조건부 그래디언트를 산출하기 위해 아래의 식 2를 이용하여 수행될 수 있다.

<0019> [식 2] $\frac{1}{\rho_i} \nabla_{\boldsymbol{x}} \log p(l|\boldsymbol{x})$

<0020> ρ_i: 업데이트 단계 i에서의 스케줄링 계수

<0021> x: 현재 단계에서의 연속 변수(원자 좌표)

<0022> l: 목표 속성(예: 결합 친화도, 합성 가능성 및 선택성)

<0023> $\nabla_x \log p(l|x)$: 연속 변수 x에 대한 목표 속성의 로그우도(log-likelihood)의 기울기

<0024> 또한, 본 발명의 실시예에 따른 상기 조건부 그래디언트 업데이트부는, 상기 조건부 그래디언트를 이용하여 상기 연속 변수에 대한 파라미터를 아래 식 3을 이용하여 더 업데이트할 수 있다.

<0025> [식 3] $$\theta_i^x = \frac{\alpha_i}{\rho_i} \cdot y + \frac{\rho_{i-1}}{\rho_i} \cdot \theta_{i-1}^x + \frac{1}{\rho_i} \nabla_x \log p(l|x)$$

<0026> θ_i^x: 단계 i에서 업데이트된 연속 변수(원자 좌표)에 대한 잠재 파라미터

<0027> θ_{i-1}^x: 이전 단계$(i-1)$의 연속 변수 잠재 파라미터로, 업데이트의 기준 상태

<0028> y: 무조건적 생성(Unconditional Generation)을 위한 기본 로그릿(logit) 입력 항

<0029> α_i: 단계 i에서의 업데이트 강도를 조절하는 계수

<0030> ρ_i, ρ_{i-1}: 확산 기반 업데이트에서 사용되는 단계별 스케줄링 계수

<0031> 또한, 본 발명의 실시예에 따른 상기 조건부 그래디언트 업데이트부는, 상기 범주형 변수에 대한 파라미터를 아래의 식 4를 이용하여 더 업데이트할 수 있다.

<0032> [식 4] $$\theta_i^v = Softmax\left(e^y \cdot \theta_{i-1}^v \cdot e^{\nabla_{e_x} \log p(l|e_x)}\right)$$

<0033> $\boldsymbol{\theta}_{i-1}^{v}$: 이전 단계에서의 범주형 변수 잠재 파라미터

<0034> $\boldsymbol{y}$: 무조건적 생성(Unconditional Generation) 항을 구성하는 로그릿(logit) 벡터

<0035> $e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1}^{v}$: 기본적인 원자 타입 분포를 형성하는 무조건적 생성(Unconditional Generation) 항

<0036> $\nabla_{e_x} \log p(l|e_x)$: 범주형 변수가 목표 속성 l에 기여하는 정도를 나타내는 조건부 그래디언트

<0037> $e^{\nabla_{e_x} \log p(l|e_x)}$: 조건부 그래디언트 항을 확률적 업데이트에 반영하기 위한 스케일링 인자

<0038> $Softmax(\cdot)$: 입력된 로그릿(logit)들을 확률적 연속 표현으로 정규화하는 함수

<0039> 또한, 본 발명의 실시예에 따른 상기 목표 속성은 단백질-리간드 결합 친화도(binding affinity), 리간드의 합성 가능성(synthetic feasibility) 및 선택성(selectivity)을 포함하는 약물 설계의 중요한 화학적 및 구조적 특성일 수 있다.

<0040> 또한, 본 발명의 실시예에 따른 상기 3차원 분자 구조 생성부는, 상기 업데이트된 파라미터에 기초하여 상기 단백질 구조 및 상기 목표 속성에 조건화된 상기 3차원 분자 구조를 생성하기 위하여 아래의 식 5를 이용하여 수행될 수 있다.

<0041> [식 5] $p_{\phi}(\boldsymbol{m}|p,l) = \int p_{\phi}(\boldsymbol{m}|\boldsymbol{\theta}_n, p, l) p(\boldsymbol{\theta}_0) \prod_{i=1}^{n} p_U(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1},\ p, l; \alpha_i)\, d\boldsymbol{\theta}_{1:n}$

<0042> $p(\boldsymbol{\theta}_0)$: 초기 파라미터 분포

<0043> $p_U(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1}, p, l; \alpha_i)$: 단계 i의 베이즈 업데이트 분포

<0044> $p_\phi(\boldsymbol{m}|\boldsymbol{\theta}_n, p, l)$: 파라미터 $\boldsymbol{\theta}_n$에 조건화된 분자 구조 생성 확률 분포

<0045> 또한, 본 발명의 실시예에 따른 상기 제어형 베이즈 흐름 신경망 모델부는, 샘플 공간이 아닌 매개변수 공간에서 상기 그래디언트를 이용하여 상기 파라미터를 업데이트하도록 구성될 수 있다.

<0046> 상기 기술적 과제를 달성하기 위하여, 제어형 그래디언트-베이즈 흐름 네트워크 기반 3차원 분자 구조 생성 시스템이 3차원 분자 구조를 생성하는 방법에 있어서, (A) 분자를 구성하는 연속 변수 및 범주형 변수를 수신하는 단계; (B) 수신된 상기 연속 변수 및 상기 범주형 변수를 제어형 베이즈 흐름 신경망 모델부에서 이용할 수 있는 형태로 전처리하는 단계; (C) 전처리된 상기 연속 변수 및 상기 범주형 변수에 기초하여 베이즈 흐름(Bayesian Flow)과 그래디언트 기반 업데이트를 수행하여, 원자 좌표 및 원자 타입을 산출함으로써 3차원 분자 구조를 생성하는 단계; 및 (D) 상기 생성된 3차원 분자 구조의 결과를 출력하는 단계; 를 제공한다.

<0047> 또한, 본 발명의 실시예에 따른 상기 (C) 단계는, (a) 상기 범주형 변수를 그래디언트 업데이트에 활용할 수 있도록 연속 공간으로 변환하는 단계; (b) 상기 연속 변수 및 상기 연속 공간으로 변환된 범주형 변수에 기초하여 산출된 조건부 그래디언트를 이용하여 파라미터를 업데이트하는 단계; 및 (c) 업데이트된 상기 파라미터에 기초하여, 단백질 구조 및 목표 속성에 조건화된 3차원 분자 구조를 생성

하는 단계; 를 포함할 수 있다.

<0048> 또한, 본 발명의 실시예에 따른 상기 (a) 단계는, 원자 타입 형태의 상기 범주형 변수를 아래의 식6에 의해 연속 공간으로 변환하도록 구성될 수 있다.

<0049> [식 6] $h(\boldsymbol{\theta}_{i-1}, \boldsymbol{y}, \alpha) = Softmax(e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1})$

<0050> $\boldsymbol{\theta}_{i-1}$: 이전 단계에서의 잠재 파라미터

<0051> $\boldsymbol{y}$: 범주형 변수(예: 원자 타입)를 변환하기 위한 입력 벡터

<0052> α: 해당 단계에서의 베이즈 업데이트 계수

<0053> $e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1}$: 범주형 변수 입력과 잠재 파라미터의 결합 결과

<0054> $Softmax(\cdot)$: 입력 벡터를 확률적 연속 표현으로 변환하는 함수

<0055> 또한, 본 발명의 실시예에 따른 상기 (b) 단계는, 단백질 구조 및 목표 속성에 기초하여 상기 조건부 그래디언트를 산출하기 위해 아래의 식 7을 이용하여 수행될 수 있다.

<0056> [식 7] $\frac{1}{\rho_i} \nabla_{\boldsymbol{x}} \log p(l|\boldsymbol{x})$

<0057> ρ_i: 업데이트 단계 i에서의 스케줄링 계수

<0058> $\boldsymbol{x}$: 현재 단계에서의 연속 변수(원자 좌표)

<0059> l: 목표 속성(예: 결합 친화도, 합성 가능성 및 선택성)

<0060> $\nabla_{\boldsymbol{x}} \log p(l|\boldsymbol{x})$: 연속 변수 $\boldsymbol{x}$에 대한 목표 속성의 로그우도(log-likelihood)의 기울기

<0061> 또한, 본 발명의 실시예에 따른 상기 (b) 단계는, 상기 조건부 그래디언트를

이용하여 상기 연속 변수에 대한 파라미터를 아래 식 8을 이용하여 더 업데이트할 수 있다.

<0062> [식 8] $\boldsymbol{\theta}_i^x = \frac{\alpha_i}{\rho_i} \cdot \boldsymbol{y} + \frac{\rho_{i-1}}{\rho_i} \cdot \boldsymbol{\theta}_{i-1}^x + \frac{1}{\rho_i} \nabla_x \log p(l|\boldsymbol{x})$

<0063> $\boldsymbol{\theta}_i^x$: 단계 i에서 업데이트된 연속 변수(원자 좌표)에 대한 잠재 파라미터

<0064> $\boldsymbol{\theta}_{i-1}^x$: 이전 단계$(i-1)$의 연속 변수 잠재 파라미터로, 업데이트의 기준 상태

<0065> $\boldsymbol{y}$: 무조건적 생성(Unconditional Generation)을 위한 기본 로그릿(logit) 입력 항

<0066> α_i: 단계 i에서의 업데이트 강도를 조절하는 계수

<0067> ρ_i, ρ_{i-1}: 확산 기반 업데이트에서 사용되는 단계별 스케줄링 계수

<0068> 또한, 본 발명의 실시예에 따른 상기 (b) 단계는, 상기 범주형 변수에 대한 파라미터를 아래의 식 9를 이용하여 더 업데이트할 수 있다.

<0069> [식 9] $\boldsymbol{\theta}_i^v = Softmax\left(e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1}^v \cdot e^{\nabla_{e_x} \log p(l|e_x)}\right)$

<0070> $\boldsymbol{\theta}_{i-1}^v$: 이전 단계에서의 범주형 변수 잠재 파라미터

<0071> $\boldsymbol{y}$: 무조건적 생성(Unconditional Generation) 항을 구성하는 로그릿(logit) 벡터

<0072> $e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1}^v$: 기본적인 원자 타입 분포를 형성하는 무조건적 생성(Unconditional Generation) 항

<0073> $\nabla_{e_x} \log p(l|e_x)$: 범주형 변수가 목표 속성 l에 기여하는 정도를 나타내는 조건

부 그래디언트

<0074> $e^{\nabla_{e_x} \log p(l|e_x)}$: 조건부 그래디언트 항을 확률적 업데이트에 반영하기 위한 스케일링 인자

<0075> $Softmax(\cdot)$: 입력된 로그릿(logit)들을 확률적 연속 표현으로 정규화하는 함수

<0076> 또한, 본 발명의 실시예에 따른 상기 목표 속성은 단백질-리간드 결합 친화도(binding affinity), 리간드의 합성 가능성(synthetic feasibility) 및 선택성(selectivity)을 포함하는 약물 설계의 중요한 화학적 및 구조적 특성일 수 있다.

<0077> 또한, 본 발명의 실시예에 따른 상기 (c) 단계는, 상기 업데이트된 파라미터에 기초하여 상기 단백질 구조 및 상기 목표 속성에 조건화된 상기 3차원 분자 구조를 생성하기 위하여 아래의 식 10을 이용하여 수행될 수 있다.

<0078> [식 10] $p_{\phi}(\boldsymbol{m}|p,l) = \int p_{\phi}(\boldsymbol{m}|\boldsymbol{\theta}_n,p,l)p(\boldsymbol{\theta}_0)\prod_{i=1}^{n} p_U(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1},\ p,l;\alpha_i)\,d\boldsymbol{\theta}_{1:n}$

<0079> $p(\boldsymbol{\theta}_0)$: 초기 파라미터 분포

<0080> $p_U(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1},\ p,l;\alpha_i)$: 단계 i의 베이즈 업데이트 분포

<0081> $p_{\phi}(\boldsymbol{m}|\boldsymbol{\theta}_n,p,l)$: 파라미터 $\boldsymbol{\theta}_n$에 조건화된 분자 구조 생성 확률 분포

<0082> 또한, 본 발명의 실시예에 따른 상기 (C) 단계는, 샘플 공간이 아닌 매개변수 공간에서 상기 그래디언트를 이용하여 상기 파라미터를 업데이트하도록 구성될 수 있다.

<0083> 상기 기술적 과제를 달성하기 위하여, 컴퓨터가 읽기 가능한 프로그램에 있어서, 3차원 분자 구조를 생성하는 방법을 수행하는 컴퓨터가 읽기 가능한 프로그램을 제공할 수 있다.

【발명의 효과】

<0084> 본 발명의 일 실시예에 따르면, 연속적인 원자 좌표 변수와 범주형 원자 타입 변수 간의 상호작용을 매개변수 공간에서 통합적으로 반영함으로써, 생성 과정에서의 화학적 맥락 손실을 방지하고 보다 정합성 높은 3차원 분자 구조를 얻을 수 있다.

<0085> 또한, 본 발명의 일 실시예에 따르면 베이즈 업데이트를 통해 범주형 변수를 연속 표현으로 변환함으로써 그래디언트 기반 조건부 유도의 안정성이 향상되고, 확산 기반 모델에서 나타나던 비효율적·불안정한 유도 신호 문제를 개선할 수 있다.

<0086> 또한, 본 발명의 일 실시예에 따르면 샘플 공간이 아닌 매개변수 공간에서 그래디언트를 이용한 파라미터 업데이트가 수행되므로, 3차원 분자의 화학적 및 구조적 유효성을 유지한 상태에서 목표 속성에 대한 안정적인 제어가 가능하고, 구조 붕괴 없이 고품질의 분자 구조를 산출할 수 있다.

<0087> 본 발명의 효과는 상기한 효과로 한정되는 것은 아니며, 본 발명의 설명 또는 청구범위에 기재된 발명의 구성으로부터 추론 가능한 모든 효과를 포함하는 것으로 이해되어야 한다.

【도면의 간단한 설명】

<0088> 도 1은 본 발명의 실시예에 따른 제어형 그래디언트-베이즈 흐름 네트워크 기반 3차원 분자 구조 생성 시스템의 개요도를 도시한 도면이다.

도 2는 본 발명의 실시예에 따른 제어형 그래디언트-베이즈 흐름 네트워크 기반 3차원 분자 구조 생성 시스템의 전체 구성을 도시한 블록도이다.

도 3은 본 발명의 실시예에 따른 제어형 베이즈 흐름 신경망 모델부의 세부 구성을 도시한 블록도이다.

도 4는 본 발명의 실시예에 따른 제어형 그래디언트-베이즈 흐름 네트워크 기반 3차원 분자 구조 생성 시스템의 전체 흐름을 도시한 도면이다.

도 5는 본 발명의 실시예에 따른 제어형 베이즈 흐름 신경망 모델 내부의 세부 절차를 도시한 흐름도이다.

도 6은 본 발명의 실시예에 따른 CBYG 프레임워크와 종래의 확산 기반 모델의 그래디언트 안정성 비교를 도시한 도면이다.

도 7은 본 발명의 실시예에 따른 CBYG 기반 생성 과정과 종래의 확산 기반 생성 과정 간의 생성 동작 차이를 비교한 도면이다.

도 8은 본 발명의 실시예에 따른 4F1M 단백질 포켓에서의 모델별 분자 생성 결과를 비교하여 도시한 도면이다.

도 9는 본 발명의 실시예에 따른 제어형 그래디언트-베이즈 흐름 네트워크 기반 3차원 분자 구조 생성 시스템을 구현하는 컴퓨팅 장치를 도시한다.

【발명을 실시하기 위한 구체적인 내용】

<0089> 이하에서는 첨부한 도면을 참조하여 본 발명을 설명하기로 한다. 그러나 본

발명은 여러 가지 상이한 형태로 구현될 수 있으며, 따라서 여기에서 설명하는 실시예로 한정되는 것은 아니다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

<0090> 명세서 전체에서, 어떤 부분이 다른 부분과 "연결(접속, 접촉, 결합)"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라, 그 중간에 다른 부재를 사이에 두고 "간접적으로 연결"되어 있는 경우도 포함한다. 또한 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 구비할 수 있다는 것을 의미한다.

<0091> 본 명세서에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

<0092> 본 명세서에서, "모듈"은 하드웨어, 소프트웨어 또는 펌웨어로 구성된 유닛을 포함하며, 예컨대 로직, 논리 블록, 부품, 또는 회로 등의 용어와 상호 호환적으로 사용될 수 있다. 모듈은 일체로 구성된 부품 또는 하나 또는 그 이상의 기능

을 수행하는 최소 단위 또는 그 일부가 될 수 있다. 예컨대 모듈은 ASIC(application-specific integrated circuit)으로 구성될 수 있다.

<0093> 이하 첨부된 도면을 참고하여 본 발명의 실시예를 상세히 설명하기로 한다.

<0094>

<0095> 도 1은 본 발명의 실시예에 따른 제어형 그래디언트-베이즈 흐름 네트워크 기반 3차원 분자 구조 생성 시스템 (100)의 개요도를 도시한 도면이다.

<0096> 제어형 그래디언트-베이즈 흐름 네트워크 기반 3차원 분자 구조 생성 시스템(100)은 좌측의 연속 변수 및 범주형 변수를 입력 변수로 활용하여 3차원 분자 구조를 산출할 수 있다.

<0097> 본 발명에서의 상기 그래디언트(guidance gradient)는 단백질 구조와 목표 속성에 기반하여, 분자 구조가 해당 목표를 잘 만족하는 방향으로 파라미터를 업데이트하도록 유도하는 조건부 기울기 정보를 의미한다.

<0098> 상기 연속 변수는 분자를 구성하는 각 원자의 3차원 위치를 나타내는 원자 좌표(x, y, z) 정보로서, 실수(real number)로 표현되는 연속적 데이터이며 분자의 공간적 형태와 구조적 배치를 산출하는 핵심 요소이다. 반면, 상기 범주형 변수는 원자의 종류(C, N, O, S 등)를 나타내는 이산적 데이터로서 화학적 성질과 반응성을 산출하는 중요한 특성이 있으나, 이산적 특정으로 인해 직접적인 그래디언트 적용이 불가능한 한계를 가진다. 따라서, 상기 범주형 변수에 베이즈 흐름을 작용하여 연속 공간 표현으로 변환함으로써 상기 연속 변수와 함께 그래디언트 기반 최적화를 적용할 수 있도록 하여 3차원 분자 구조 산출 과정에서 연속 변수 및 범주형

변수의 상호작용을 일관되고 안정적으로 반영할 수 있다.

<0099> 도 2는 본 발명의 실시예에 따른 제어형 그래디언트-베이즈 흐름 네트워크 기반 3차원 분자 구조 생성 시스템 (100)의 전체 구성을 도시한 블록도이다.

<0100> 제어형 그래디언트-베이즈 흐름 네트워크 기반 3차원 분자 구조 생성 시스템(100)은 입력 변수 수신부(110), 데이터 전처리부(120), 제어형 베이즈 흐름 신경망 모델부(130) 및 결과 출력부(140)을 포함할 수 있다.

<0101> 입력 변수 수신부(110)는 분자를 구성하는 연속 변수 및 범주형 변수를 외부로부터 수신할 수 있다. 상기 연속 변수는 각 원자의 3차원 좌표(x, y, z)를 포함하는 실수(real number) 기반 데이터로, 분자의 기하적 구조와 공간적 배치를 정의한다. 또한, 상기 범주형 변수는 원자의 종류(C, N, O, S 등)를 나타내는 이산적 데이터이며, 분자의 전하 분포, 반응성 및 결합 특성과 같은 화학적 성질을 산출한다. 이와 같이 상기 연속 변수와 상기 범주형 변수는 데이터의 형식과 특징이 서로 상이하므로, 각각 적절한 방식으로 처리될 필요가 있다.

<0102> 데이터 전처리부(120)는 입력 변수 수신부(110)에 수신된 상기 연속 변수 및 상기 범주형 변수를 제어형 베이즈 흐름 신경망 모델부(130)에서 이용할 수 있는 형태로 제공하기 위한 전처리를 수행할 수 있다.

<0103> 상기 연속 변수가 원자 좌표(x, y, z)로 구성되는 경우에는 좌표의 정규화나 중심 이동과 같은 공간적 정렬 과정이 포함될 수 있으며, 상기 범주형 변수가 원자 타입(C, N, O, S 등)과 같이 이산적 데이터인 경우에는 원-핫 인코딩(one-hot encoding) 또는 임베딩(embedding) 벡터로의 변환과 같은 표현 변환 과정이 포함될

수 있다. 또한, 분자마다 원자 수가 상이할 수 있으므로 입력 크기를 통일하기 위한 패딩(padding)이나 유효 원자와 패딩을 구별하기 위한 마스크(mask) 생성 과정이 포함될 수 있다.

<0104> 이와 같은 전처리 과정들은 이후 과정에서 제어형 베이즈 흐름 신경망 모델이 연속 변수 및 범주형 변수를 안정적이고 일관되게 처리할 수 있도록 하기 위한 과정이다.

<0105> 상기 제어형 베이즈 흐름 신경망 모델은 연속 변수 및 범주형 변수에 대한 베이즈 기반 변환과 조건부 그래디언트 업데이트를 통하여 단백질 구조 및 목표 속성에 최적화된 3차원 분자 구조를 생성하도록 제어되는 신경망 모델이다.

<0106> 또한, 본 발명에서의 상기 그래디언트(guidance gradient)는 단백질 구조와 목표 속성에 기반하여, 분자 구조가 해당 목표를 잘 만족하는 방향으로 파라미터(parameter)를 업데이트하도록 유도하는 조건부 기울기 정보를 의미한다.

<0107> 여기서, 본 발명에서의 상기 파라미터는 3차원 분자 구조를 구성하기 위한 잠재적 표현(latent representation)으로서, 상기 연속 변수 및 상기 범주형 변수가 매개변수 공간(parameter space)에서 표현되는 내부 변수 집합을 의미한다. 상기 파라미터는 실제 분자의 좌표나 원자 타입과 같은 샘플 공간(sample space)의 값을 직접 조작하거나 갱신하는 것이 아니라, 생성 과정 전반에서 그래디언트를 이용하여 연속적이고 안정적으로 갱신되도록 정의된 모델 내부의 잠재적 변수들이다. 따라서 상기 파라미터의 업데이트는 구조 붕괴를 유발할 수 있는 직접적 좌표 이동 방식과 달리, 목표 속성에 부합하는 방향으로 분자 구조를 점진적·안정적으로 최

적화하기 위한 핵심 절차를 의미한다.

<0108> 상기 목표 속성은 단백질-리간드 결합 친화도(binding affinity), 리간드의 합성 가능성(synthetic feasibility)및 선택성(selectivity) 등을 포함하는 약물 설계의 중요한 화학적 및 구조적 특성이다.

<0109> 제어형 베이즈 흐름 신경망 모델부(130)는 전처리된 상기 연속 변수 및 상기 범주형 변수에 대해 베이즈 흐름(Bayesian Flow)을 통해 상기 범주형 변수를 연속 공간 표현으로 변환한 뒤, 상기 연속 변수 및 연속 공간으로 변환된 상기 범주형 변수에 그래디언트 기반으로 업데이트를 수행하여, 상기 연속 변수 및 상기 범주형 변수에 대응하는 3차원 분자의 원자 좌표 및 원자 타입을 산출함으로써 3차원 분자 구조를 생성하여 산출한다.

<0110> 이를 통해 종래의 확산 기반 생성 모델의 한계였던 이산적 변수의 미분 불가능성 문제를 극복하고, 결합친화도, 선택성 및 합성가능성 등 목표 속성에 부합하는 방향으로 분자 구조를 점진적으로 최적화할 수 있다. 더 나아가 제어형 베이즈 흐름 신경망 모델부(130)는 샘플 공간이 아닌 매개변수 공간(parameter space)에서 그래디언트를 이용하여 파라미터를 업데이트하므로, 생성 과정의 안정성과 연속성이 크게 향상된다.

<0111> 상기 샘플 공간은 분자의 원자 좌표 및 원자 타입 등과 같이 분자의 최종 구조가 존재하는 공간을 의미하며, 상기 샘플 공간에서 직접 변수를 조작하는 경우 미세한 좌표 변동에도 구조적 붕괴가 발생하기 쉽다. 반면, 상기 매개변수 공간은 분자를 구성하기 위한 연속적 표현 또는 잠재 파라미터가 존재하는 공간으로, 상기

범주형 변수가 베이즈 흐름에 의해 연속 공간 표현으로 변환된 후 그래디언트를 이용하여 안정적으로 업데이트할 수 있다. 상기 매개변수 공간에서의 업데이트를 통해 일관된 구조적 형태를 갖는 3차원 분자 구조 생성이 가능하다.

<0112> 결과 출력부(140)는 제어형 베이즈 흐름 신경망 모델부(130)에서 산출된 3차원 분자 구조를 외부로 출력할 수 있다. 결과 출력부(140)는 원자 좌표 및 원자 타입 정보 등을 포함하는 3차원 분자 구조 데이터를 후속 분석, 시각화 또는 외부 시스템과의 연계를 위하여 제공될 수 있다. 또한, 결과 출력부(140)는 제어형 베이즈 흐름 신경망 모델의 산출 결과가 안정적으로 전달될 수 있도록 해당 구조 데이터를 지정된 형식으로 정리하여 제공할 수 있다.

<0113> 도 3은 본 발명의 실시예에 따른 제어형 베이즈 흐름 신경망 모델부(130)의 세부 구성을 도시한 블록도이다.

<0114> 상기 제어형 베이즈 흐름 신경망 모델은 연속 변수 및 범주형 변수에 대한 베이즈 기반 변환과 조건부 그래디언트 업데이트를 통하여 단백질 구조 및 목표 속성에 최적화된 3차원 분자 구조를 생성하도록 제어되는 신경망 모델이다.

<0115> 또한, 본 발명에서의 상기 그래디언트(guidance gradient)는 단백질 구조와 목표 속성에 기반하여, 분자 구조가 해당 목표를 잘 만족하는 방향으로 파라미터를 업데이트하도록 유도하는 조건부 기울기 정보를 의미한다.

<0116> 상기 목표 속성은 단백질-리간드 결합 친화도(binding affinity), 리간드의 합성 가능성(synthetic feasibility)및 선택성(selectivity) 등을 포함하는 약물 설계의 중요한 화학적 및 구조적 특성이다.

<0117> 제어형 베이즈 흐름 신경망 모델부(130)는 베이즈 기반 범주형 변수 변환부(131), 조건부 그래디언트 업데이트부(132) 및 3차원 분자 구조 생성부(133)를 포함할 수 있다.

<0118> 본 발명은 제어형 베이즈 흐름 신경망(Control guided Bayesian Flow Network, BFN)을 이용하여 3차원 분자 구조를 생성할 수 있다.

<0119> 상기 제어형 베이즈 흐름 신경망은 다음 수학식과 같은 베이즈 생성 모델로 구현될 수 있다.

<0120> 【수학식 1】

$$p_{\phi}(\boldsymbol{m}) = \int p_{\phi}(\boldsymbol{m}|\boldsymbol{\theta}_n)p(\boldsymbol{\theta}_0)\prod_{i=1}^{n} p_U(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1};\alpha_i)\,d\boldsymbol{\theta}_{1:n}$$

<0121> $\boldsymbol{\theta}_0$: 모델의 초기 파라미터

<0122> $p_U(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1};\alpha_i)$: 단계 i에서의 베이즈 업데이트 분포

<0123> α_i: 해당 단계에서의 업데이트 강도 및 정밀도를 조절하는 계수

<0124> $\boldsymbol{\theta}_n$: 일련의 업데이트 과정을 거쳐 도출된 최종 파라미터

<0125> $p_{\phi}(\boldsymbol{m}|\boldsymbol{\theta}_n)$: 최종 파라미터에 조건화된 분자 구조 생성 확률분포

<0126> $d\boldsymbol{\theta}_{1:n}$: 전체 업데이트 경로에 대한 적분

<0127> 을 각각 의미하며, 초기 파라미터 $\boldsymbol{\theta}_0$가 각 단계의 업데이트 분포 $p_U(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1};\alpha_i)$를 거쳐 최종 파라미터 $\boldsymbol{\theta}_n$가 산출되고, 상기 최종 파라미터 $\boldsymbol{\theta}_n$에 기초하여 분자 구조 $\boldsymbol{m}$을 생성할 수 있다.

<0128> 베이즈 기반 범주형 변수 변환부(131)는 원자 타입 형태의 범주형 변수(categorical variable)를 그래디언트 업데이트에 활용할 수 있도록 연속 공간(latent continuous space)으로 변환할 수 있다.

<0129> 상기 범주형 변수의 연속 공간 표현으로의 변환은 베이즈 업데이트 기반의 변환 함수로 정의될 수 있으며, 다음의 수학식과 같다.

<0130> 【수학식 2】

$$h(\boldsymbol{\theta}_{i-1}, \boldsymbol{y}, \alpha) = Softmax(e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1})$$

<0131> $\boldsymbol{\theta}_{i-1}$: 이전 단계에서의 잠재 파라미터

<0132> $\boldsymbol{y}$: 범주형 변수(예: 원자 타입)를 변환하기 위한 입력 벡터

<0133> α: 해당 단계에서의 베이즈 업데이트 계수

<0134> $e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1}$: 범주형 변수 입력과 잠재 파라미터의 결합 결과

<0135> $Softmax(\cdot)$: 입력 벡터를 확률적 연속 표현으로 변환하는 함수

<0136> 수학식 2는 상기 범주형 변수 $\boldsymbol{y}$와 이전 단계 잠재 파라미터 $\boldsymbol{\theta}_{i-1}$의 결합 결과를 $Softmax$ 함수에 적용하여, 이산적으로 표현된 상기 범주형 변수를 연속 공간에서의 확률적 벡터 표현으로 변환하기 위한 함수이다. 이를 통해 범주형 변수 또는 이후 단계에서 그래디언트 기반 업데이트가 가능하도록 연속적 형태로 매핑될 수 있다.

<0137> 조건부 그래디언트 업데이트부(132)는 전처리된 연속 변수 및 베이즈 기반 범주형 변수 변환부(131)를 통해 연속 공간으로 변환된 범주형 변수에 대하여, 단

백질 구조 정보 및 목표 속성에 기초하여 조건부 그래디언트를 산출하고, 상기 산출된 그래디언트를 이용하여 잠재 파라미터를 업데이트할 수 있다. 상기 조건부 그래디언트의 산출은 다음 수학식으로 정의될 수 있다.

<0138> 【수학식 3】

$$\frac{1}{\rho_i}\nabla_x \log p(l|x)$$

<0139> ρ_i: 업데이트 단계 i에서의 스케줄링 계수

<0140> x: 현재 단계에서의 연속 변수(원자 좌표)

<0141> l: 목표 속성(예: 결합 친화도, 합성 가능성 및 선택성)

<0142> $\nabla_x \log p(l|x)$: 연속 변수 x에 대한 목표 속성의 로그우도(log-likelihood)의 기울기

<0143> 수학식 3은 목표 속성 l이 주어졌을 때, 현재 원자 좌표 x가 상기 목표 속성에 기여하는 방향성을 나타내는 조건부 그래디언트 항을 정의한다.

<0144> 즉, 상기 그래디언트 항은 단백질 포켓 구조 및 상기 목표 속성에 기초하여 잠재 파라미터가 '목표 속성을 높이는 방향'으로 업데이트 되도록 제어할 수 있다.

<0145> 또한, 업데이트 계수 $\frac{1}{\rho_i}$는 단계별로 적용되는 스케줄링 인자로, 업데이트 과정에서 상기 그래디언트의 크기가 안정적으로 조절되도록 한다.

<0146> 이어서, 수학식 3을 사용하는 본 발명의 실시예에 따른 연속 변수의 조건부 그래디언트 업데이트는 다음 수학식으로 구현될 수 있다.

<0147> 【수학식 4】

$$\boldsymbol{\theta}_i^{\boldsymbol{x}} = \frac{\alpha_i}{\rho_i} \cdot \boldsymbol{y} + \frac{\rho_{i-1}}{\rho_i} \cdot \boldsymbol{\theta}_{i-1}^{\boldsymbol{x}} + \frac{1}{\rho_i} \nabla_{\boldsymbol{x}} \log p(l|\boldsymbol{x})$$

<0148> $\boldsymbol{\theta}_i^{\boldsymbol{x}}$: 단계 i에서 업데이트된 연속 변수(원자 좌표)에 대한 잠재 파라미터

<0149> $\boldsymbol{\theta}_{i-1}^{\boldsymbol{x}}$: 이전 단계$(i-1)$의 연속 변수 잠재 파라미터로, 업데이트의 기준 상태

<0150> $\boldsymbol{y}$: 무조건적 생성(Unconditional Generation)을 위한 기본 로그릿(logit) 입력 항

<0151> α_i: 단계 i에서의 업데이트 강도를 조절하는 계수

<0152> ρ_i, ρ_{i-1}: 확산 기반 업데이트에서 사용되는 단계별 스케줄링 계수로, 업데이트 항들의 크기를 안정적으로 조절하기 위한 비율 조정에 사용됨

<0153> 수학식 4는 상기 연속 변수를 단계적으로 업데이트하기 위한 실시예로서, 첫번째 항 $\frac{\alpha_i}{\rho_i} \cdot \boldsymbol{y}$는 무조건적 생성 방향을 나타내며, 두번째 항 $\frac{\rho_{i-1}}{\rho_i} \cdot \boldsymbol{\theta}_{i-1}^{\boldsymbol{x}}$는 이전 상태를 반영하는 안정화 항이며, 세번째 항 $\frac{1}{\rho_i} \nabla_{\boldsymbol{x}} \log p(l|\boldsymbol{x})$는 목표 속성 기반 조건부 그래디언트 항을 의미한다.

<0154> 세 항이 결합됨으로써 상기 연속 변수는 목표 속성에 부합하는 방향으로 점진적으로 최적화될 수 있다.

<0155> 한편, 상기 범주형 변수에 대해서도 상기 목표 속성에 기초하여 산출된 조건부 그래디언트를 기반으로 업데이트될 수 있으며, 상기 범주형 변수에 대한 업데이트는 다음 수학식으로 정의될 수 있다.

<0156> 【수학식 5】

$$\boldsymbol{\theta}_i^v = Softmax\left(e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1}^v \cdot e^{\nabla_{e_x} \log p(l|e_x)}\right)$$

<0157> $\boldsymbol{\theta}_{i-1}^v$: 이전 단계에서의 범주형 변수 잠재 파라미터

<0158> $\boldsymbol{y}$: 무조건적 생성(Unconditional Generation) 항을 구성하는 로그릿(logit) 벡터

<0159> $e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1}^v$: 기본적인 원자 타입 분포를 형성하는 무조건적 생성(Unconditional Generation) 항

<0160> $\nabla_{e_x} \log p(l|e_x)$: 범주형 변수가 목표 속성 l에 기여하는 정도를 나타내는 조건부 그래디언트

<0161> $e^{\nabla_{e_x} \log p(l|e_x)}$: 조건부 그래디언트 항을 확률적 업데이트에 반영하기 위한 스케일링 인자

<0162> $Softmax(\cdot)$: 입력된 로그릿(logit)들을 확률적 연속 표현으로 정규화하는 함수

<0163> 따라서, 수학식 5는 상기 목표 속성(예: 결합 친화도(binding affinity), 합성 가능성(synthetic feasibility) 및 선택성(selectivity) 등)에 대한 조건부 그래디언트를 반영함으로써, 베이즈 기반 범주형 변수 변환부(131)를 통해 연속 공간으로 변환된 원자 타입 분포를 상기 목표 속성에 유리한 방향으로 최적화하기 위한 업데이트 과정을 정의한다.

<0164> 3차원 분자 구조 생성부(133)는 조건부 그래디언트 업데이트부(132)를 통해

단계적으로 업데이트된 최종 파라미터 $\boldsymbol{\theta}_n$에 기초하여, 단백질 구조 p 및 목표 속성 l에 조건화된 3차원 분자 구조를 생성할 수 있다.

<0165> 상기 3차원 분자 구조 생성은 다음의 수학식을 통해 수행될 수 있다.

<0166> 【수학식 6】

$$p_{\phi}(\boldsymbol{m}|p,l) = \int p_{\phi}(\boldsymbol{m}|\boldsymbol{\theta}_n,p,l)p(\boldsymbol{\theta}_0)\prod_{i=1}^{n} p_U(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1},\ p,l;\alpha_i)\,d\boldsymbol{\theta}_{1:n}$$

<0167> $p(\boldsymbol{\theta}_0)$: 초기 파라미터 분포

<0168> $p_U(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1},\ p,l;\alpha_i)$: 단계 i의 베이즈 업데이트 분포

<0169> $p_{\phi}(\boldsymbol{m}|\boldsymbol{\theta}_n,p,l)$: 최종 파라미터 $\boldsymbol{\theta}_n$에 조건화된 분자 구조 생성 확률 분포

<0170> 따라서, 3차원 분자 구조 생성부(133)는 베이즈 업데이트를 통해 도출된 최종 파라미터에 기초하여, 단백질 포켓 구조 및 목표 속성에 부합하는 3차원 분자 구조 $\boldsymbol{m}$을 생성할 수 있다.

<0171> 도 4는 본 발명의 실시예에 따른 제어형 그래디언트-베이즈 흐름 네트워크 기반 3차원 분자 구조 생성 시스템(100)의 전체 흐름을 도시한 도면이다.

<0172> 단계(S110)에서는 분자를 구성하는 연속 변수 및 범주형 변수를 외부로부터 수신한다. 상기 연속 변수는 각 원자의 3차원 좌표(x, y, z)를 포함하는 실수(real number) 기반 데이터로, 분자의 기하적 구조와 공간적 배치를 정의한다. 또한, 상기 범주형 변수는 원자의 종류(C, N, O, S 등)를 나타내는 이산적 데이터이며, 분자의 전하 분포, 반응성 및 결합 특성과 같은 화학적 성질을 산출한다. 이와 같이 상기 연속 변수와 상기 범주형 변수는 데이터의 형식과 특징이 서로 상이하므로,

각각 적절한 방식으로 처리될 필요가 있다.

<0173> 단계(S120)에서는 전술된 단계(S110)에 수신된 상기 연속 변수 및 상기 범주형 변수를 제어형 베이즈 흐름 신경망 모델에서 이용할 수 있는 형태로 제공하기 위한 전처리를 수행한다.

<0174> 상기 연속 변수가 원자 좌표(x, y, z)로 구성되는 경우에는 좌표의 정규화나 중심 이동과 같은 공간적 정렬 과정이 포함될 수 있으며, 상기 범주형 변수가 원자 타입(C, N, O, S 등)과 같이 이산적 데이터인 경우에는 원-핫 인코딩(one-hot encoding) 또는 임베딩(embedding) 벡터로의 변환과 같은 표현 변환 과정이 포함될 수 있다. 또한, 분자마다 원자 수가 상이할 수 있으므로 입력 크기를 통일하기 위한 패딩(padding)이나 유효 원자와 패딩을 구별하기 위한 마스크(mask) 생성 과정이 포함될 수 있다.

<0175> 이와 같은 전처리 과정들은 이후 과정에서 상기 제어형 베이즈 흐름 신경망 모델이 연속 변수 및 범주형 변수를 안정적이고 일관되게 처리할 수 있도록 하기 위한 과정이다.

<0176> 상기 제어형 베이즈 흐름 신경망 모델은 연속 변수 및 범주형 변수에 대한 베이즈 기반 변환과 조건부 그래디언트 업데이트를 통하여 단백질 구조 및 목표 속성에 최적화된 3차원 분자 구조를 생성하도록 제어되는 신경망 모델이다.

<0177> 또한, 본 발명에서의 상기 그래디언트(guidance gradient)는 단백질 구조와 목표 속성에 기반하여, 분자 구조가 해당 목표를 잘 만족하는 방향으로 파라미터를 업데이트하도록 유도하는 조건부 기울기 정보를 의미한다.

<0178> 여기서, 본 발명에서의 상기 파라미터는 3차원 분자 구조를 구성하기 위한 잠재적 표현(latent representation)으로서, 상기 연속 변수 및 상기 범주형 변수가 매개변수 공간(parameter space)에서 표현되는 내부 변수 집합을 의미한다. 상기 파라미터는 실제 분자의 좌표나 원자 타입과 같은 샘플 공간(sample space)의 값을 직접 조작하거나 갱신하는 것이 아니라, 생성 과정 전반에서 그래디언트를 이용하여 연속적이고 안정적으로 갱신되도록 정의된 모델 내부의 잠재적 변수들이다. 따라서 상기 파라미터의 업데이트는 구조 붕괴를 유발할 수 있는 직접적 좌표 이동 방식과 달리, 목표 속성에 부합하는 방향으로 분자 구조를 점진적·안정적으로 최적화하기 위한 핵심 절차를 의미한다.

<0179> 단계(S130)에서는 전처리된 상기 연속 변수 및 상기 범주형 변수에 대해 베이즈 흐름(Bayesian Flow)을 통해 상기 범주형 변수를 연속 공간 표현으로 변환한 뒤, 상기 연속 변수 및 연속 공간으로 변환된 상기 범주형 변수에 그래디언트 기반으로 업데이트를 수행하여, 상기 연속 변수 및 상기 범주형 변수에 대응하는 3차원 분자의 원자 좌표 및 원자 타입을 산출함으로써 3차원 분자 구조를 생성하여 산출한다.

<0180> 이를 통해 종래의 확산 기반 생성 모델의 한계였던 이산적 변수의 미분 불가능성 문제를 극복하고, 결합친화도, 선택성 및 합성가능성 등 목표 속성에 부합하는 방향으로 분자 구조를 점진적으로 최적화할 수 있다. 더 나아가 제어형 베이즈 흐름 신경망 모델은 샘플 공간이 아닌 매개변수 공간(parameter space)에서 그래디언트를 이용하여 파라미터를 업데이트하므로, 생성 과정의 안정성과 연속성이 크게

향상된다.

<0181> 상기 샘플 공간은 분자의 원자 좌표 및 원자 타입 등과 같이 분자의 최종 구조가 존재하는 공간을 의미하며, 상기 샘플 공간에서 직접 변수를 조작하는 경우 미세한 좌표 변동에도 구조적 붕괴가 발생하기 쉽다. 반면, 상기 매개변수 공간은 분자를 구성하기 위한 연속적 표현 또는 잠재 파라미터가 존재하는 공간으로, 상기 범주형 변수가 베이즈 흐름에 의해 연속 공간 표현으로 변환된 후 그래디언트를 이용하여 안정적으로 업데이트할 수 있다. 상기 매개변수 공간에서의 업데이트를 통해 일관된 구조적 형태를 갖는 3차원 분자 구조 생성이 가능하다.

<0182> 단계(S140)에서는 전술된 단계(S130)에서 산출된 3차원 분자 구조를 외부로 출력한다. 단계(S140)에서는 원자 좌표 및 원자 타입 정보 등을 포함하는 3차원 분자 구조 데이터를 후속 분석, 시각화 또는 외부 시스템과의 연계를 위하여 제공될 수 있다. 또한, 제어형 베이즈 흐름 신경망 모델의 산출 결과가 안정적으로 전달될 수 있도록 해당 구조 데이터를 지정된 형식으로 정리하여 제공할 수 있다.

<0183> 도 5는 본 발명의 실시예에 따른 제어형 베이즈 흐름 신경망 모델 내부의 세부 절차를 도시한 흐름도이다.

<0184> 상기 제어형 베이즈 흐름 신경망 모델은 연속 변수 및 범주형 변수에 대한 베이즈 기반 변환과 조건부 그래디언트 업데이트를 통하여 단백질 구조 및 목표 속성에 최적화된 3차원 분자 구조를 생성하도록 제어되는 신경망 모델이다.

<0185> 또한, 본 발명에서의 상기 그래디언트(guidance gradient)는 단백질 구조와 목표 속성에 기반하여, 분자 구조가 해당 목표를 잘 만족하는 방향으로 파라미터를

업데이트하도록 유도하는 조건부 기울기 정보를 의미한다.

<0186> 상기 목표 속성은 단백질-리간드 결합 친화도(binding affinity), 리간드의 합성 가능성(synthetic feasibility)및 선택성(selectivity) 등을 포함하는 약물 설계의 중요한 화학적 및 구조적 특성이다.

<0187> 본 발명은 제어형 베이즈 흐름 신경망(Control guided Bayesian Flow Network, BFN)을 이용하여 3차원 분자 구조를 생성한다.

<0188> 상기 제어형 베이즈 흐름 신경망은 다음 수학식과 같은 베이즈 생성 모델로 구현된다.

<0189> 【수학식 7】

$$p_{\phi}(\boldsymbol{m}) = \int p_{\phi}(\boldsymbol{m}|\boldsymbol{\theta}_n)p(\boldsymbol{\theta}_0)\prod_{i=1}^{n} p_U(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1};\alpha_i)\, d\boldsymbol{\theta}_{1:n}$$

<0190> $\boldsymbol{\theta}_0$: 모델의 초기 파라미터

<0191> $p_U(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1};\alpha_i)$: 단계 i에서의 베이즈 업데이트 분포

<0192> α_i: 해당 단계에서의 업데이트 강도 및 정밀도를 조절하는 계수

<0193> $\boldsymbol{\theta}_n$: 일련의 업데이트 과정을 거쳐 도출된 최종 파라미터

<0194> $p_{\phi}(\boldsymbol{m}|\boldsymbol{\theta}_n)$: 최종 파라미터에 조건화된 분자 구조 생성 확률분포

<0195> $d\boldsymbol{\theta}_{1:n}$: 전체 업데이트 경로에 대한 적분

<0196> 을 각각 의미하며, 초기 파라미터 $\boldsymbol{\theta}_0$가 각 단계의 업데이트 분포 $p_U(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1};\alpha_i)$를 거쳐 최종 파라미터 $\boldsymbol{\theta}_n$가 산출되고, 상기 최종 파라미터 $\boldsymbol{\theta}_n$에 기

초하여 분자 구조 $\boldsymbol{m}$을 생성할 수 있다.

<0197> 단계(S131)에서는 원자 타입 형태의 범주형 변수(categorical variable)를 그래디언트 업데이트에 활용할 수 있도록 연속 공간(latent continuous space)으로 변환한다.

<0198> 상기 범주형 변수의 연속 공간 표현으로의 변환은 베이즈 업데이트 기반의 변환 함수로 정의될 수 있으며, 다음의 수학식과 같다.

<0199> 【수학식 8】

$$h(\boldsymbol{\theta}_{i-1}, \boldsymbol{y}, \alpha) = Softmax(e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1})$$

<0200> $\boldsymbol{\theta}_{i-1}$: 이전 단계에서의 잠재 파라미터

<0201> $\boldsymbol{y}$: 범주형 변수(예: 원자 타입)를 변환하기 위한 입력 벡터

<0202> α: 해당 단계에서의 베이즈 업데이트 계수

<0203> $e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1}$: 범주형 변수 입력과 잠재 파라미터의 결합 결과

<0204> $Softmax(\cdot)$: 입력 벡터를 확률적 연속 표현으로 변환하는 함수

<0205> 수학식 8은 상기 범주형 변수 $\boldsymbol{y}$와 이전 단계 잠재 파라미터 $\boldsymbol{\theta}_{i-1}$의 결합 결과를 $Softmax$ 함수에 적용하여, 이산적으로 표현된 상기 범주형 변수를 연속 공간에서의 확률적 벡터 표현으로 변환하기 위한 함수이다. 이를 통해 범주형 변수 또는 이후 단계에서 그래디언트 기반 업데이트가 가능하도록 연속적 형태로 매핑될 수 있다.

<0206> 단계(S132)에서는 전처리된 연속 변수 및 연속 공간으로 변환된 범주형 변수

에 대하여, 단백질 구조 정보 및 목표 속성에 기초하여 조건부 그래디언트를 산출하고, 상기 산출된 그래디언트를 이용하여 잠재 파라미터를 업데이트할 수 있다. 상기 조건부 그래디언트의 산출은 다음 수학식으로 정의될 수 있다.

<0207> 【수학식 9】

$$\frac{1}{\rho_i}\nabla_{\boldsymbol{x}} \log p(l|\boldsymbol{x})$$

<0208> ρ_i: 업데이트 단계 i에서의 스케줄링 계수

<0209> $\boldsymbol{x}$: 현재 단계에서의 연속 변수(원자 좌표)

<0210> l: 목표 속성(예: 결합 친화도, 합성 가능성 및 선택성)

<0211> $\nabla_{\boldsymbol{x}} \log p(l|\boldsymbol{x})$: 연속 변수 $\boldsymbol{x}$에 대한 목표 속성의 로그우도(log-likelihood)의 기울기

<0212> 수학식 9는 목표 속성 l이 주어졌을 때, 현재 원자 좌표 $\boldsymbol{x}$가 상기 목표 속성에 기여하는 방향성을 나타내는 조건부 그래디언트 항을 정의한다.

<0213> 즉, 상기 그래디언트 항은 단백질 포켓 구조 및 상기 목표 속성에 기초하여 잠재 파라미터가 '목표 속성을 높이는 방향'으로 업데이트 되도록 제어할 수 있다.

<0214> 또한, 업데이트 계수 $\frac{1}{\rho_i}$는 단계별로 적용되는 스케줄링 인자로, 업데이트 과정에서 상기 그래디언트의 크기가 안정적으로 조절되도록 한다.

<0215> 이어서, 수학식 9을 사용하는 본 발명의 실시예에 따른 연속 변수의 조건부 그래디언트 업데이트는 다음 수학식으로 구현된다.

<0216> 【수학식 10】

$$\boldsymbol{\theta}_i^{\boldsymbol{x}} = \frac{\alpha_i}{\rho_i} \cdot \boldsymbol{y} + \frac{\rho_{i-1}}{\rho_i} \cdot \boldsymbol{\theta}_{i-1}^{\boldsymbol{x}} + \frac{1}{\rho_i} \nabla_{\boldsymbol{x}} \log p(l|\boldsymbol{x})$$

<0217> $\boldsymbol{\theta}_i^{\boldsymbol{x}}$: 단계 i에서 업데이트된 연속 변수(원자 좌표)에 대한 잠재 파라미터

<0218> $\boldsymbol{\theta}_{i-1}^{\boldsymbol{x}}$: 이전 단계$(i-1)$의 연속 변수 잠재 파라미터로, 업데이트의 기준 상태

<0219> $\boldsymbol{y}$: 무조건적 생성(Unconditional Generation)을 위한 기본 로그릿(logit) 입력 항

<0220> α_i: 단계 i에서의 업데이트 강도를 조절하는 계수

<0221> ρ_i, ρ_{i-1}: 확산 기반 업데이트에서 사용되는 단계별 스케줄링 계수로, 업데이트 항들의 크기를 안정적으로 조절하기 위한 비율 조정에 사용됨

<0222> 수학식 10은 상기 연속 변수를 단계적으로 업데이트하기 위한 실시예로서, 첫번째 항 $\frac{\alpha_i}{\rho_i} \cdot \boldsymbol{y}$는 무조건적 생성 방향을 나타내며, 두번째 항 $\frac{\rho_{i-1}}{\rho_i} \cdot \boldsymbol{\theta}_{i-1}^{\boldsymbol{x}}$는 이전 상태를 반영하는 안정화 항이며, 세번째 항 $\frac{1}{\rho_i} \nabla_{\boldsymbol{x}} \log p(l|\boldsymbol{x})$는 목표 속성 기반 조건부 그래디언트 항을 의미한다.

<0223> 세 항이 결합됨으로써 상기 연속 변수는 목표 속성에 부합하는 방향으로 점진적으로 최적화될 수 있다.

<0224> 한편, 상기 범주형 변수에 대해서도 상기 목표 속성에 기초하여 산출된 조건부 그래디언트를 기반으로 업데이트될 수 있으며, 상기 범주형 변수에 대한 업데이트는 다음 수학식으로 정의된다.

<0225> 【수학식 11】

$$\boldsymbol{\theta}_i^v = Softmax\left(e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1}^v \cdot e^{\nabla_{e_x} \log p(l|e_x)}\right)$$

<0226> $\boldsymbol{\theta}_{i-1}^v$: 이전 단계에서의 범주형 변수 잠재 파라미터

<0227> $\boldsymbol{y}$: 무조건적 생성(Unconditional Generation) 항을 구성하는 로그릿(logit) 벡터

<0228> $e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1}^v$: 기본적인 원자 타입 분포를 형성하는 무조건적 생성(Unconditional Generation) 항

<0229> $\nabla_{e_x} \log p(l|e_x)$: 범주형 변수가 목표 속성 l에 기여하는 정도를 나타내는 조건부 그래디언트

<0230> $e^{\nabla_{e_x} \log p(l|e_x)}$: 조건부 그래디언트 항을 확률적 업데이트에 반영하기 위한 스케일링 인자

<0231> $Softmax(\cdot)$: 입력된 로그릿(logit)들을 확률적 연속 표현으로 정규화하는 함수

<0232> 따라서, 수학식 11은 상기 목표 속성(예: 결합 친화도(binding affinity), 합성 가능성(synthetic feasibility) 및 선택성(selectivity) 등)에 대한 조건부 그래디언트를 반영함으로써, 베이즈 기반 범주형 변수 변환부(131)를 통해 연속 공간으로 변환된 원자 타입 분포를 상기 목표 속성에 유리한 방향으로 최적화하기 위한 업데이트 과정을 정의한다.

<0233> 단계(S133)에서는 전술된 단계(S132)를 통해 단계적으로 업데이트된 최종 파

라미터 $\boldsymbol{\theta}_n$에 기초하여, 단백질 구조 p 및 목표 속성 l에 조건화된 3차원 분자 구조를 생성한다.

<0234> 상기 3차원 분자 구조 생성은 다음의 수학식을 통해 수행된다.

<0235> 【수학식 12】

$$p_{\phi}(\boldsymbol{m}|p,l) = \int p_{\phi}(\boldsymbol{m}|\boldsymbol{\theta}_n,p,l)p(\boldsymbol{\theta}_0)\prod_{i=1}^{n} p_U(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1},\ p,l;\alpha_i)\,d\boldsymbol{\theta}_{1:n}$$

<0236> $p(\boldsymbol{\theta}_0)$: 초기 파라미터 분포

<0237> $p_U(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1},\ p,l;\alpha_i)$: 단계 i의 베이즈 업데이트 분포

<0238> $p_{\phi}(\boldsymbol{m}|\boldsymbol{\theta}_n,p,l)$: 최종 파라미터 $\boldsymbol{\theta}_n$에 조건화된 분자 구조 생성 확률 분포

<0239> 따라서, 단계(S133)에서는 베이즈 업데이트를 통해 도출된 최종 파라미터에 기초하여, 단백질 포켓 구조 및 목표 속성에 부합하는 3차원 분자 구조 $\boldsymbol{m}$을 생성할 수 있다.

<0240> 도 6은 본 발명의 실시예에 따른 CBYG 프레임워크와 종래의 확산 기반 모델의 그래디언트 안정성 비교를 도시한 도면이다.

<0241> 도 6은 도시된 바와 같이, 본 발명의 제안된 CBYG(Controllable Bayesian Flow Network with Guidance) 프레임워크에서 사용되는 그래디언트의 동작 특성과 종래의 확산 기반 모델(diffusion model)에서 사용되는 그래디언트의 동작 특성을 비교한다.

<0242> 본 발명의 제안된 CBYG(Controllable Bayesian Flow Network with Guidance)

프레임워크는 베이즈 흐름(Bayesian Flow) 기반 업데이트와 조건부 그래디언트(guidance gradient)를 결합하여, 연속 변수(원자 좌표) 및 범주형 변수(원자 타입)를 동시에 안정적으로 최적화함으로써 단백질 구조에 적합한 3차원 분자 구조를 생성하는 제어형 분자 생성 모델 프레임워크이다.

<0243> (A) CBYG 프레임워크의 그래디언트는 매개변수 공간(parameter space) 기반 베이즈 흐름 업데이트와 조건부 그래디언트 업데이트에 의해, 생성 단계 전반에 걸쳐 매끄럽고 연속적인 경로를 유지하면서 안정적으로 최적화가 진행되는 모습을 나타낸다. (A) CBYG 프레임워크의 그래디언트가 목표 속성에 부합하는 방향으로 일관되게 유도되고 있음을 의미한다.

<0244> 반면, (B) 종래의 확산 기반 모델의 그래디언트는 샘플 공간(sample space)에서 직접 그래디언트를 업데이트 함으로써, 특정 단계에서 불연속적 이동, 분산되는 경로 및 급격한 변화가 발생하는 모습을 보여준다. 이러한 불안정성은 샘플의 작은 교란에도 구조 붕괴가 발생하기 쉬운 (B) 종래의 확산 기반 모델의 한계를 반영한다.

<0245> 따라서, 본 발명의 제안된 CBYG 프레임워크는 종래 기법과 비교하여 그래디언트의 안정성, 연속성 및 목표 속성에 대한 수렴 효율이 우수함을 도 6을 통해 확인할 수 있다.

<0246> 도 7은 본 발명의 실시예에 따른 CBYG 기반 생성 과정과 종래의 확산 기반 생성 과정 간의 생성 동작 차이를 비교한 도면이다.

<0247> (C) 종래 확산 기반 생성 과정은 샘플 공간(sample space)에서 분자 구조가

직접 샘플링되며, 원자 좌표 및 원자 타입의 확률 분포가 불안정하게 변동함에 따라 생성 경로가 붕괴되거나 목표 속성 반영이 제한되는 문제점을 나타낸다.

<0248> 반면, (D) CBYG 기반 생성 과정에서는 매개변수 공간(parameter space)에서 연속 변수 및 변환된 범주형 변수가 안정적으로 업데이트되며, 조건부 그래디언트가 목표 속성에 따라 분자 구조를 일관된 방향으로 최적화하도록 유도함으로써 생성 과정의 안정성과 정확성이 향상된다.

<0249> 또한, 도 7에 도시된 3차원 그래프의 x축 및 y축은 생성 과정에서 샘플링된 분자들을 2차원 임베딩 공간에 투영한 좌표값을 나타내고, z축은 훈련 데이터 및 모델이 학습한 확률 분포 또는 밀도 값에 해당하는 높이 값을 나타낸다.

<0250> 따라서, x, y, z 값은 실제 분자의 공간 좌표계를 의미하는 것이 아니라, 생성 경로와 분포 변화를 시각적으로 표현하기 위한 모델 내부 표현(latent visualization)임을 알 수 있다.

<0251> 또한, 종래 확산 기반 모델이 샘플 공간에서의 직접 샘플링으로 인해 분자 구조가 불안정하게 생성되는 반면, 본 발명의 CBYG 기반 생성 과정은 매개변수 공간에서의 안정적인 업데이트와 조건부 그래디언트 기반 제어를 통해 목표 속성(예: 결합 친화도, 합성 가능성 및 선택성)에 부합하는 3차원 분자 구조를 안정적으로 생성할 수 있음을 시각적으로 나타낸다.

<0252> 도 8은 본 발명의 실시예에 따른 4F1M 단백질 포켓에서의 모델별 분자 생성 결과를 비교하여 도시한 도면이다.

<0253> 우선 좌측의 4F1M 단백질 포켓에 대한 레퍼런스(reference) 리간드 구조와

함께, 종래의 ALIDiff, DecompOpt 및 MolCRAFT 모델이 생성한 분자 구조를 순차적으로 나타내고, 마지막으로 본 발명의 CBYG 모델이 생성한 분자 구조를 도시한다.

<0254> 종래 ALIDiff, DecompOpt 및 MolCRAFT 모델이 생성한 분자들은 레퍼런스 구조 대비 포켓 내부에서의 배치가 불안정하거나, 결합 자세(binding pose)에서 일부 원자 그룹이 포켓 외부로 이탈하는 등 결합 친화도 및 구조적 적합성이 낮은 형태를 나타내는 경우가 확인된다.

<0255> 반면, 우측에 도시된 본 발명의 상기 CBYG 모델은 단백질 포켓의 공간적 제약 및 목표 속성(예: 결합 친화도)에 부합하는 방향으로 최적화된 분자 구조를 생성하며, 레퍼런스와 유사한 결합 자세를 유지하면서 포켓 내부에 안정적으로 배치된 구조를 나타낸다. 특히, 상기 CBYG 모델은 원자 간 거리, 방향성 및 결합 골격이 보다 일관된 형태로 유지되며, 생성된 분자가 단백질 포켓 내에서 높은 구조적 적합성을 가지는 것이 확인된다.

<0256> 따라서, 도 8은 본 발명의 상기 CBYG 모델이 종래의 확산 기반 모델들에 비해 단백질-리간드 상호작용을 더 정확하게 반영하고, 구조적 안정성과 목표 속성 적합성이 우수한 3차원 분자 구조를 생성할 수 있음을 시각적으로 입증한다.

<0257> 도 9는 본 발명의 실시예에 따른 제어형 그래디언트-베이즈 흐름 네트워크 기반 3차원 분자 구조 생성 시스템을 구현하는 컴퓨팅 장치를 도시한다.

<0258> 도 4~ 도 8에 의해 설명된 본 발명의 실시에는 적어도 하나의 프로세서에 의해 동작하는 컴퓨팅 장치(900)로 구현될 수 있다.

<0259> 컴퓨팅 장치(900)는 프로세서(910), 메모리(920), 스토리지(930), 통신 인터

페이스(940), 시스템 인터커넥트(950) 및 디스플레이(960)을 포함할 수 있다.

<0260> 프로세서(910)은, CPU(Central Processing Unit), MPU(Micro Processor Unit), MCU(Micro Controller Unit), GPU(Graphic Processing Unit) 및 APU(Application Processing Unit)을 포함한다.

<0261> 메모리(920)은 프로세서(910)와 상호작용하여 프로그램이 효율적으로 실행될 수 있도록 데이터를 저장하고 필요한 정보에 빠르게 접근할 수 있도록 하는 기능을 수행한다. 메모리(920)은 레지스터, 캐시 메모리, 주 메모리, 읽기 전용 메모리, 가상 메모리, 비휘발성 메모리 중 적어도 하나를 포함한다.

<0262> 스토리지(930)은 데이터를 영구적으로 저장하고 관리하는 역할을 한다. 스토리지는 컴퓨팅 시스템이 꺼지거나 재부팅된 후에도 데이터를 보존하며, 운영 체제, 애플리케이션, 사용자 파일 등을 저장하는 데 사용된다. 스토리지(930)은, 하드 디스크 드라이브(HDD), 솔리드 스테이트 드라이브(SSD), 광학 디스크, 네트워크 스토리지 및 클라우드 스토리지 중 적어도 하나를 포함한다.

<0263> 통신 인터페이스(940)는 컴퓨팅 시스템 내부 및 외부의 다양한 장치들 간에 데이터를 주고받기 위한 경로를 제공한다. 통신 인터페이스(940)는 USB(Universal Serial Bus), PCIe(Peripheral Component Interconnect Express), SATA(Serial ATA), Ethernet, Wi-Fi, Thunderbolt 및 HDMI(High-Definition Multimedia Interface) 중 적어도 하나의 통신 방식을 지원할 수 있다.

<0264> 시스템 인터커넥트(950)는 컴퓨팅 시스템 내부에서 다양한 구성 요소들 간의 데이터와 신호를 주고받는 역할을 한다. 시스템 인터커넥트(950)는, 버스(Bus), 포

인트-투-포인트(Point-to-Point), 크로스바 스위치(Crossbar Switch), 네트워크-온-칩(Network-on-Chip, NoC) 중 적어도 하나의 방식을 지원할 수 있다.

<0265> 디스플레이(960)은 컴퓨팅 시스템의 출력 장치로서, 사용자에게 시각적인 정보를 제공하는 기능을 수행한다.

<0266> 전술한 구성에 의하여, 본 발명의 실시예에 따른 프로그램은, 프로세서(910)에 의해 실행되는 명령어들에 기초하여 실행되며, 메모리(920) 또는 스토리지(930)에 저장될 수 있다.

<0267> 전술한 본 발명의 실시예에 따른 방법은 다양한 컴퓨터 구성요소를 통하여 실행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능한 기록매체에 기록될 수 있다. 컴퓨터 판독 가능한 기록매체는 프로그램 명령어, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 컴퓨터 판독 가능한 기록매체에 기록되는 프로그램 명령은 본 발명의 실시예를 위하여 특별히 설계되고 구성된 것이거나, 컴퓨터 소프트웨어 분야의 통상의 기술자에게 공지되어 사용가능한 것일 수 있다. 컴퓨터 판독 가능한 기록매체는, 하드디스크, 플로피디스크, 자기테이프 등의 자기기록 매체, CD-ROM, DVD 등의 광기록 매체, 플롭티컬디스크 등의 자기-광 매체, ROM, RAM, 플래시 메모리 등과 같이, 프로그램 명령을 저장하고 수행하도록 구성된 하드웨어를 포함한다. 프로그램 명령은, 컴파일러에 의해 만들어지는 기계어 코드, 인터프리터를 사용하여 컴퓨터에서 실행될 수 있는 고급언어 코드를 포함한다. 하드웨어는 본 발명에 따른 방법을 처리하기 위하여 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있고, 그 역도 마찬가지이다.

<0268> 본 발명의 실시예에 따른 방법은 프로그램 명령 형태로 전자장치에서 실행될 수 있다. 전자장치는 스마트폰이나 스마트패드 등의 휴대용 통신 장치, 컴퓨터 장치, 휴대용 멀티미디어 장치, 휴대용 의료 기기, 카메라, 웨어러블 장치, 가전 장치를 포함한다.

<0269> 본 발명의 실시예에 따른 방법은 컴퓨터 프로그램 제품에 포함되어 제공될 수 있다. 컴퓨터 프로그램 제품은 상품으로서 판매자 및 구매자 간에 거래될 수 있다. 컴퓨터 프로그램 제품은 기기로 읽을 수 있는 기록매체의 형태로, 또는 어플리케이션 스토어를 통해 온라인으로 배포될 수 있다. 온라인 배포의 경우에, 컴퓨터 프로그램 제품의 적어도 일부는 제조사의 서버, 어플리케이션 스토어의 서버, 또는 중계 서버의 메모리와 같은 저장 매체에 적어도 일시 저장되거나, 임시적으로 생성될 수 있다.

<0270> 본 발명의 실시예에 따른 구성요소, 예컨대 모듈 또는 프로그램 각각은 단수 또는 복수의 서브 구성요소로 구성될 수 있으며, 이러한 서브 구성요소들 중 일부 서브 구성요소가 생략되거나, 또는 다른 서브 구성요소가 더 포함될 수 있다. 일부 구성요소들(모듈 또는 프로그램)은 하나의 개체로 통합되어, 통합되기 이전의 각각의 해당 구성요소에 의해 수행되는 기능을 동일 또는 유사하게 수행할 수 있다. 본 발명의 실시예에 따른 모듈, 프로그램 또는 다른 구성요소에 의해 수행되는 동작들은 순차적, 병렬적, 반복적 또는 휴리스틱하게 실행되거나, 적어도 일부 동작이 다른 순서로 실행되거나, 생략되거나, 또는 다른 동작이 추가될 수 있다.

<0271> 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의

통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다.

<0272> 본 발명의 범위는 후술하는 청구범위에 의하여 나타내어지며, 청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

【부호의 설명】

<0273> 100: 3차원 분자 구조 생성 시스템

110: 입력 변수 수신부

120: 데이터 전처리부

130: 제어형 베이즈 흐름 신경망 모델부

140: 결과 출력부

【청구범위】

【청구항 1】

제어형 그래디언트-베이즈 흐름 네트워크 기반 3차원 분자 구조 생성 시스템에 있어서,

분자를 구성하는 연속 변수 및 범주형 변수를 수신하는 입력 변수 수신부;

수신된 상기 연속 변수 및 상기 범주형 변수를 제어형 베이즈 흐름 신경망 모델부에서 이용할 수 있는 형태로 전처리하는 데이터 전처리부;

전처리된 상기 연속 변수 및 상기 범주형 변수에 기초하여 베이즈 흐름(Bayesian Flow)과 그래디언트 기반 업데이트를 수행하여, 원자 좌표 및 원자 타입을 산출함으로써 3차원 분자 구조를 생성하는 상기 제어형 베이즈 흐름 신경망 모델부; 및

상기 생성된 3차원 분자 구조의 결과를 출력하는 결과 출력부;

를 포함하는 것인 3차원 분자 구조 생성 시스템.

【청구항 2】

제1항에 있어서,

상기 제어형 베이즈 흐름 신경망 모델부는,

상기 범주형 변수를 그래디언트 업데이트에 활용할 수 있도록 연속 공간으로 변환하는 베이즈 기반 범주형 변수 변환부;

상기 연속 변수 및 상기 연속 공간으로 변환된 범주형 변수에 기초하여 산출

된 조건부 그래디언트를 이용하여 파라미터를 업데이트하는 조건부 그래디언트 업데이트부; 및

업데이트된 상기 파라미터에 기초하여, 단백질 구조 및 목표 속성에 조건화된 3차원 분자 구조를 생성하는 3차원 분자 구조 생성부;

를 포함하는 것인 3차원 분자 구조 생성 시스템.

【청구항 3】

제2항에 있어서,

상기 베이즈 기반 범주형 변수 변화부는,

원자 타입 형태의 상기 범주형 변수를 아래의 식1에 의해 연속 공간으로 변환하도록 구성된 것인 3차원 분자 구조 생성 시스템.

[식 1]

$$h(\boldsymbol{\theta}_{i-1}, \boldsymbol{y}, \alpha) = Softmax(e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1})$$

$\boldsymbol{\theta}_{i-1}$: 이전 단계에서의 잠재 파라미터

$\boldsymbol{y}$: 범주형 변수(예: 원자 타입)를 변환하기 위한 입력 벡터

α: 해당 단계에서의 베이즈 업데이트 계수

$e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1}$: 범주형 변수 입력과 잠재 파라미터의 결합 결과

$Softmax(\cdot)$: 입력 벡터를 확률적 연속 표현으로 변환하는 함수

【청구항 4】

제2항에 있어서,

상기 조건부 그래디언트 업데이트부는,

단백질 구조 및 목표 속성에 기초하여 상기 조건부 그래디언트를 산출하기 위해 아래의 식 2를 이용하여 수행되는 것인 3차원 분자 구조 생성 시스템.

[식 2]

$$\frac{1}{\rho_i}\nabla_{\boldsymbol{x}} \log p(l|\boldsymbol{x})$$

ρ_i: 업데이트 단계 i에서의 스케줄링 계수

$\boldsymbol{x}$: 현재 단계에서의 연속 변수(원자 좌표)

l: 목표 속성(예: 결합 친화도, 합성 가능성 및 선택성)

$\nabla_{\boldsymbol{x}} \log p(l|\boldsymbol{x})$: 연속 변수 $\boldsymbol{x}$에 대한 목표 속성의 로그우도(log-likelihood)의 기울기

【청구항 5】

제4항에 있어서,

상기 조건부 그래디언트 업데이트부는,

상기 조건부 그래디언트를 이용하여 상기 연속 변수에 대한 파라미터를 아래 식 3을 이용하여 더 업데이트하는 것인 3차원 분자 구조 생성 시스템.

[식 3]

$$\boldsymbol{\theta}_i^{\boldsymbol{x}} = \frac{\alpha_i}{\rho_i} \cdot \boldsymbol{y} + \frac{\rho_{i-1}}{\rho_i} \cdot \boldsymbol{\theta}_{i-1}^{\boldsymbol{x}} + \frac{1}{\rho_i}\nabla_{\boldsymbol{x}} \log p(l|\boldsymbol{x})$$

$\boldsymbol{\theta}_i^x$: 단계 i에서 업데이트된 연속 변수(원자 좌표)에 대한 잠재 파라미터

$\boldsymbol{\theta}_{i-1}^x$: 이전 단계$(i-1)$의 연속 변수 잠재 파라미터로, 업데이트의 기준 상태

$\boldsymbol{y}$: 무조건적 생성(Unconditional Generation)을 위한 기본 로그릿(logit) 입력 항

α_i: 단계 i에서의 업데이트 강도를 조절하는 계수

ρ_i, ρ_{i-1}: 확산 기반 업데이트에서 사용되는 단계별 스케줄링 계수

【청구항 6】

제5항에 있어서,

상기 조건부 그래디언트 업데이트부는,

상기 범주형 변수에 대한 파라미터를 아래의 식 4를 이용하여 더 업데이트하는 것인 3차원 분자 구조 생성 시스템.

[식 4]

$$\boldsymbol{\theta}_i^v = Softmax\left(e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1}^v \cdot e^{\nabla_{e_x} \log p(l|e_x)}\right)$$

$\boldsymbol{\theta}_{i-1}^v$: 이전 단계에서의 범주형 변수 잠재 파라미터

$\boldsymbol{y}$: 무조건적 생성(Unconditional Generation) 항을 구성하는 로그릿(logit) 벡터

$e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1}^v$: 기본적인 원자 타입 분포를 형성하는 무조건적 생성(Unconditional Generation) 항

$\nabla_{e_x} \log p(l|e_x)$: 범주형 변수가 목표 속성 l에 기여하는 정도를 나타내는 조건부 그래디언트

$e^{\nabla_{e_x} \log p(l|e_x)}$: 조건부 그래디언트 항을 확률적 업데이트에 반영하기 위한 스케일링 인자

$Softmax(\cdot)$: 입력된 로그릿(logit)들을 확률적 연속 표현으로 정규화하는 함수

【청구항 7】

제4항에 있어서,

상기 목표 속성은 단백질-리간드 결합 친화도(binding affinity), 리간드의 합성 가능성(synthetic feasibility) 및 선택성(selectivity)을 포함하는 약물 설계의 중요한 화학적 및 구조적 특성인 것인 3차원 분자 구조 생성 시스템.

【청구항 8】

제2항에 있어서,

상기 3차원 분자 구조 생성부는,

상기 업데이트된 파라미터에 기초하여 상기 단백질 구조 및 상기 목표 속성에 조건화된 상기 3차원 분자 구조를 생성하기 위하여 아래의 식 5를 이용하여 수행되는 것인 3차원 분자 구조 생성 시스템.

[식 5]

$$p_{\phi}(\boldsymbol{m}|p,l) = \int p_{\phi}(\boldsymbol{m}|\boldsymbol{\theta}_n,p,l)p(\boldsymbol{\theta}_0)\prod_{i=1}^{n} p_U(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1},\ p,l;\alpha_i)\,d\boldsymbol{\theta}_{1:n}$$

$p(\boldsymbol{\theta}_0)$: 초기 파라미터 분포

$p_U(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1},\ p,l;\alpha_i)$: 단계 i의 베이즈 업데이트 분포

$p_{\phi}(\boldsymbol{m}|\boldsymbol{\theta}_n,p,l)$: 파라미터 $\boldsymbol{\theta}_n$에 조건화된 분자 구조 생성 확률 분포

【청구항 9】

제2항에 있어서,

상기 제어형 베이즈 흐름 신경망 모델부는,

샘플 공간이 아닌 매개변수 공간에서 상기 그래디언트를 이용하여 상기 파라미터를 업데이트하도록 구성된 것인 3차원 분자 구조 생성 시스템.

【청구항 10】

제어형 그래디언트-베이즈 흐름 네트워크 기반 3차원 분자 구조 생성 시스템이 3차원 분자 구조를 생성하는 방법에 있어서,

(A) 분자를 구성하는 연속 변수 및 범주형 변수를 수신하는 단계;

(B) 수신된 상기 연속 변수 및 상기 범주형 변수를 제어형 베이즈 흐름 신경망 모델부에서 이용할 수 있는 형태로 전처리하는 단계;

(C) 전처리된 상기 연속 변수 및 상기 범주형 변수에 기초하여 베이즈 흐

름(Bayesian Flow)과 그래디언트 기반 업데이트를 수행하여, 원자 좌표 및 원자 타입을 산출함으로써 3차원 분자 구조를 생성하는 단계; 및

(D) 상기 생성된 3차원 분자 구조의 결과를 출력하는 단계;

를 포함하는 것인 3차원 분자 구조를 생성하는 방법.

【청구항 11】

제10항에 있어서,

상기 (C) 단계는,

(a) 상기 범주형 변수를 그래디언트 업데이트에 활용할 수 있도록 연속 공간으로 변환하는 단계;

(b) 상기 연속 변수 및 상기 연속 공간으로 변환된 범주형 변수에 기초하여 산출된 조건부 그래디언트를 이용하여 파라미터를 업데이트하는 단계; 및

(c) 업데이트된 상기 파라미터에 기초하여, 단백질 구조 및 목표 속성에 조건화된 3차원 분자 구조를 생성하는 단계;

를 포함하는 것인 3차원 분자 구조를 생성하는 방법.

【청구항 12】

제11항에 있어서,

상기 (a) 단계는,

원자 타입 형태의 상기 범주형 변수를 아래의 식6에 의해 연속 공간으로 변

환하도록 구성된 것인 3차원 분자 구조를 생성하는 방법.

[식 6]

$$h(\boldsymbol{\theta}_{i-1}, \boldsymbol{y}, \alpha) = Softmax(e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1})$$

$\boldsymbol{\theta}_{i-1}$: 이전 단계에서의 잠재 파라미터

$\boldsymbol{y}$: 범주형 변수(예: 원자 타입)를 변환하기 위한 입력 벡터

α: 해당 단계에서의 베이즈 업데이트 계수

$e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1}$: 범주형 변수 입력과 잠재 파라미터의 결합 결과

$Softmax(\cdot)$: 입력 벡터를 확률적 연속 표현으로 변환하는 함수

【청구항 13】

제11항에 있어서,

상기 (b) 단계는,

단백질 구조 및 목표 속성에 기초하여 상기 조건부 그래디언트를 산출하기 위해 아래의 식 7을 이용하여 수행되는 것인 3차원 분자 구조를 생성하는 방법.

[식 7]

$$\frac{1}{\rho_i} \nabla_{\boldsymbol{x}} \log p(l|\boldsymbol{x})$$

ρ_i: 업데이트 단계 i에서의 스케줄링 계수

$\boldsymbol{x}$: 현재 단계에서의 연속 변수(원자 좌표)

l: 목표 속성(예: 결합 친화도, 합성 가능성 및 선택성)

$\nabla_x \log p(l|x)$: 연속 변수 x에 대한 목표 속성의 로그우도(log-likelihood)의 기울기

【청구항 14】

제13항에 있어서,

상기 (b) 단계는,

상기 조건부 그래디언트를 이용하여 상기 연속 변수에 대한 파라미터를 아래 식 8을 이용하여 더 업데이트하는 것인 3차원 분자 구조를 생성하는 방법.

[식 8]

$$\theta_i^x = \frac{\alpha_i}{\rho_i} \cdot y + \frac{\rho_{i-1}}{\rho_i} \cdot \theta_{i-1}^x + \frac{1}{\rho_i} \nabla_x \log p(l|x)$$

θ_i^x: 단계 i에서 업데이트된 연속 변수(원자 좌표)에 대한 잠재 파라미터

θ_{i-1}^x: 이전 단계$(i-1)$의 연속 변수 잠재 파라미터로, 업데이트의 기준 상태

y: 무조건적 생성(Unconditional Generation)을 위한 기본 로그릿(logit) 입력 항

α_i: 단계 i에서의 업데이트 강도를 조절하는 계수

ρ_i, ρ_{i-1}: 확산 기반 업데이트에서 사용되는 단계별 스케줄링 계수

【청구항 15】

제14항에 있어서,

상기 (b) 단계는,

상기 범주형 변수에 대한 파라미터를 아래의 식 9를 이용하여 더 업데이트하는 것인 3차원 분자 구조를 생성하는 방법.

[식 9]

$$\boldsymbol{\theta}_i^v = Softmax\left(e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1}^v \cdot e^{\nabla_{e_x} \log p(l|e_x)}\right)$$

$\boldsymbol{\theta}_{i-1}^v$: 이전 단계에서의 범주형 변수 잠재 파라미터

$\boldsymbol{y}$: 무조건적 생성(Unconditional Generation) 항을 구성하는 로그릿(logit) 벡터

$e^{\boldsymbol{y}} \cdot \boldsymbol{\theta}_{i-1}^v$: 기본적인 원자 타입 분포를 형성하는 무조건적 생성(Unconditional Generation) 항

$\nabla_{e_x} \log p(l|e_x)$: 범주형 변수가 목표 속성 l에 기여하는 정도를 나타내는 조건부 그래디언트

$e^{\nabla_{e_x} \log p(l|e_x)}$: 조건부 그래디언트 항을 확률적 업데이트에 반영하기 위한 스케일링 인자

$Softmax(\cdot)$: 입력된 로그릿(logit)들을 확률적 연속 표현으로 정규화하는 함수

【청구항 16】

제11항에 있어서,

상기 목표 속성은 단백질-리간드 결합 친화도(binding affinity), 리간드의 합성 가능성(synthetic feasibility) 및 선택성(selectivity)을 포함하는 약물 설계의 중요한 화학적 및 구조적 특성인 것인 3차원 분자 구조를 생성하는 방법.

【청구항 17】

제11항에 있어서,

상기 (c) 단계는,

상기 업데이트된 파라미터에 기초하여 상기 단백질 구조 및 상기 목표 속성에 조건화된 상기 3차원 분자 구조를 생성하기 위하여 아래의 식 10을 이용하여 수행되는 것인 3차원 분자 구조를 생성하는 방법.

[식 10]

$$p_{\phi}(\boldsymbol{m}|p,l) = \int p_{\phi}(\boldsymbol{m}|\boldsymbol{\theta}_n,p,l)p(\boldsymbol{\theta}_0)\prod_{i=1}^{n} p_U(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1},\ p,l;\alpha_i)\,d\boldsymbol{\theta}_{1:n}$$

$p(\boldsymbol{\theta}_0)$: 초기 파라미터 분포

$p_U(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1},\ p,l;\alpha_i)$: 단계 i의 베이즈 업데이트 분포

$p_{\phi}(\boldsymbol{m}|\boldsymbol{\theta}_n,p,l)$: 파라미터 $\boldsymbol{\theta}_n$에 조건화된 분자 구조 생성 확률 분포

【청구항 18】

제11항에 있어서,

상기 (C) 단계는,

샘플 공간이 아닌 매개변수 공간에서 상기 그래디언트를 이용하여 상기 파라미터를 업데이트하도록 구성된 것인 3차원 분자 구조를 생성하는 방법.

【청구항 19】

컴퓨터가 읽기 가능한 프로그램에 있어서, 제10항 내지 제18항 중 어느 한 항에 따른 방법을 수행하는 컴퓨터가 읽기 가능한 프로그램.

【요약서】

【요약】

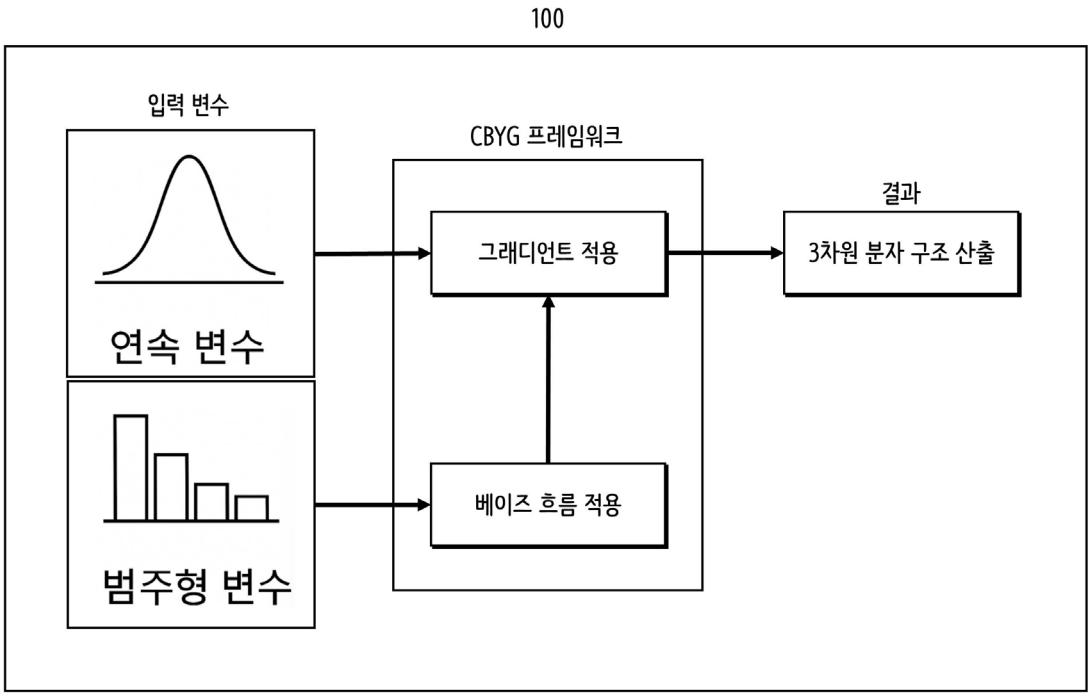
본 발명의 일실시예는 제어형 그래디언트-베이즈 흐름 네트워크 기반 3차원 분자 구조 생성 시스템에 있어서, 분자를 구성하는 연속 변수 및 범주형 변수를 수신하는 입력 변수 수신부; 수신된 상기 연속 변수 및 상기 범주형 변수를 제어형 베이즈 흐름 신경망 모델부에서 이용할 수 있는 형태로 전처리하는 데이터 전처리부; 전처리된 상기 연속 변수 및 상기 범주형 변수에 기초하여 베이즈 흐름(Bayesian Flow)과 그래디언트 기반 업데이트를 수행하여, 원자 좌표 및 원자 타입을 산출함으로써 3차원 분자 구조를 생성하는 상기 제어형 베이즈 흐름 신경망 모델부; 및 상기 생성된 3차원 분자 구조의 결과를 출력하는 결과 출력부; 를 제공한다.

【대표도】

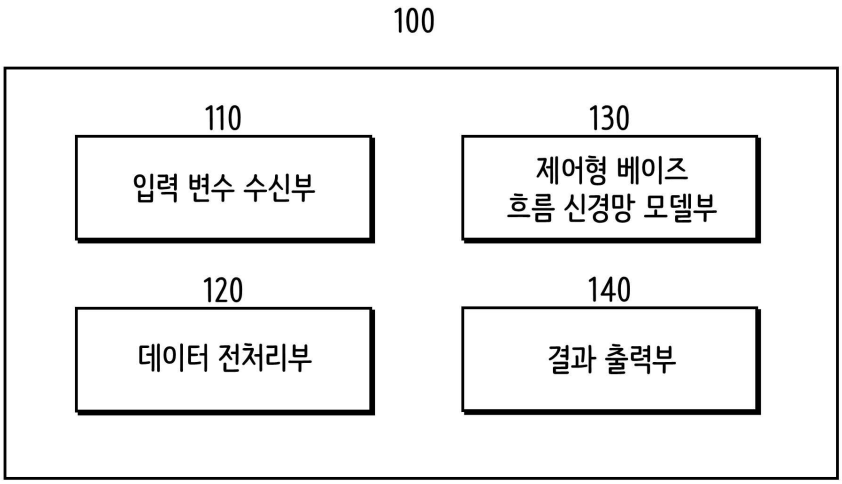
도 1

【도면】

【도 1】


100
입력 변수
연속 변수
범주형 변수
CBYG 프레임워크
그래디언트 적용
베이즈 흐름 적용
결과
3차원 분자 구조 산출

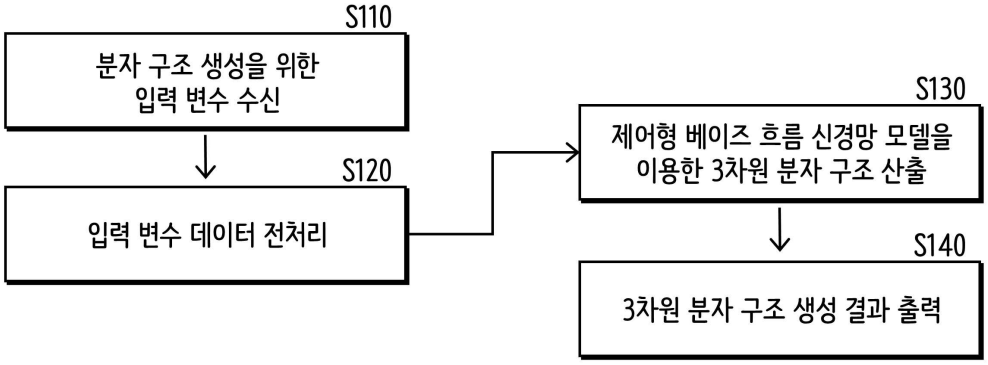
【도 2】


100
110
입력 변수 수신부
120
데이터 전처리부
130
제어형 베이즈
흐름 신경망 모델부
140
결과 출력부

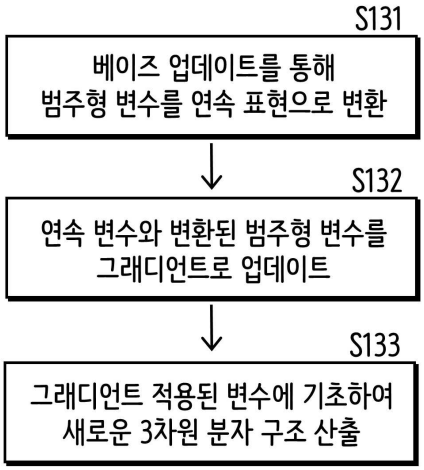
【도 3】


130
131
베이즈 기반
범주형 변수 변환부
132
조건부 그래디언트
업데이트부
133
3차원 분자
구조 생성부

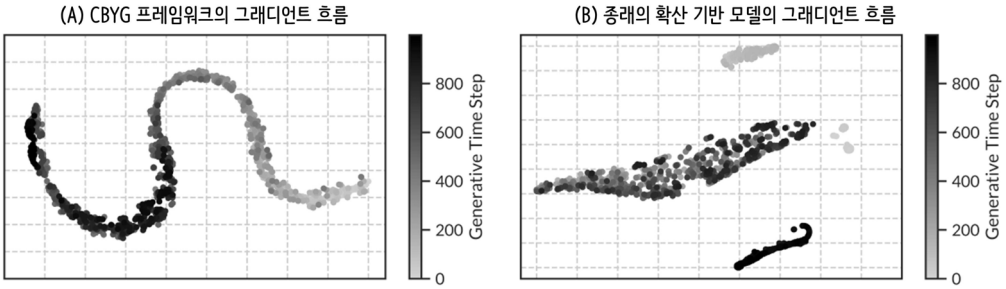
【도 4】


S110
분자 구조 생성을 위한
입력 변수 수신
S120
입력 변수 데이터 전처리
S130
제어형 베이즈 흐름 신경망 모델을
이용한 3차원 분자 구조 산출
S140
3차원 분자 구조 생성 결과 출력

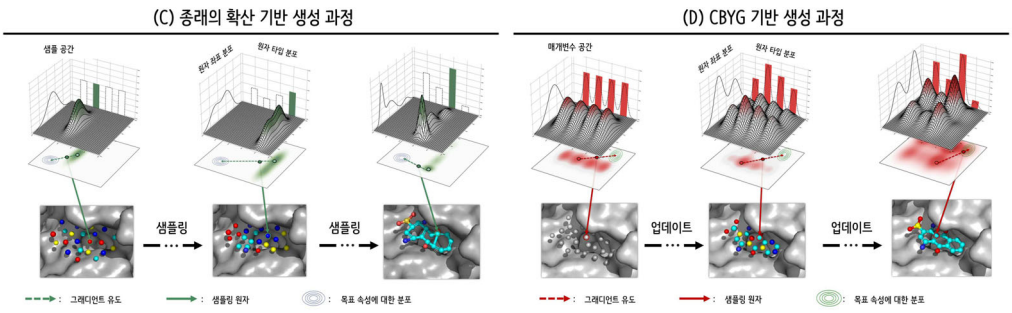
【도 5】


S131
베이즈 업데이트를 통해
범주형 변수를 연속 표현으로 변환
S132
연속 변수와 변환된 범주형 변수를
그래디언트로 업데이트
S133
그래디언트 적용된 변수에 기초하여
새로운 3차원 분자 구조 산출

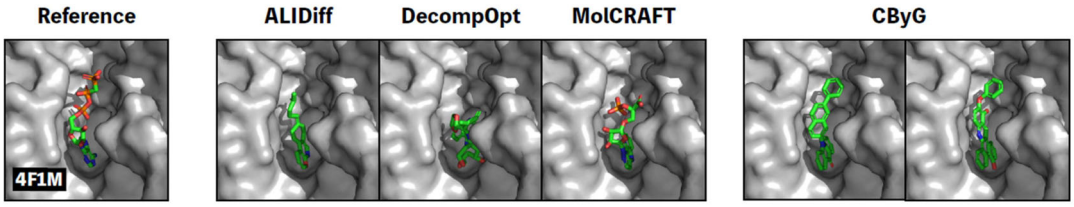
【도 6】


(A) CBYG 프레임워크의 그래디언트 흐름
800
600
400
200
0
Generative Time Step
(B) 종래의 확산 기반 모델의 그래디언트 흐름
800
600
400
200
0
Generative Time Step

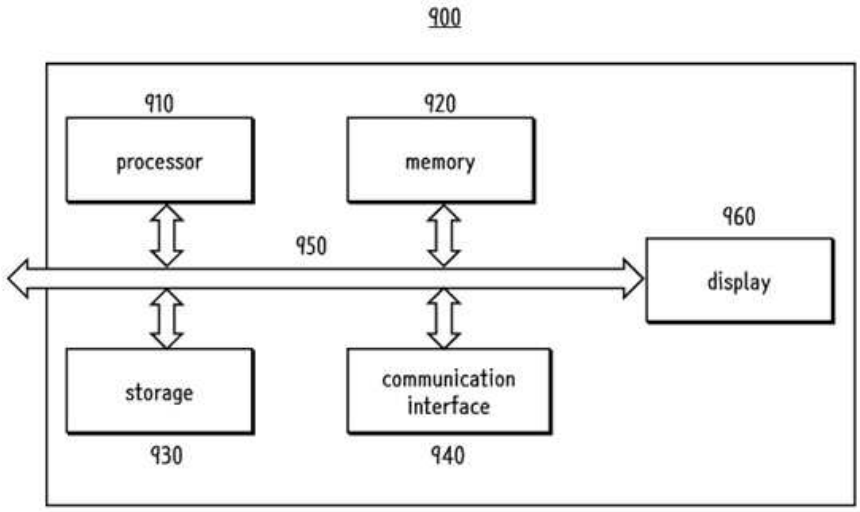
【도 7】


(C) 종래의 확산 기반 생성 과정
샘플링
샘플링
(D) CBYG 기반 생성 과정
업데이트
업데이트

【도 8】


Reference
ALIDiff
DecompOpt
MolCRAFT
CByG
4F1M

【도 9】


900
910
processor
920
memory
960
display
950
storage
930
communication interface
940