Lab Seminar

Discovering Symbolic Models from DL with Inductive Biases

MOAI Lab Internship
Park Chaewon

The Unreasonable Effectiveness of Mathematics in the Natural Science

"물리 법칙을 정식화하는 데 있어 수학이라는 언어가 이렇게까지
절묘하게 들어맞는다는 사실 —그 기적성은— 우리가 이해하지도 못하고,
감히 받을 자격조차 없는, 그야말로 경이로운 선물이다."

— Eugene Wigner

"The miracle of the appropriateness of the language of
mathematics for the formulation of the laws of physics
is a wonderful gift which we neither understand nor deserve."

The Unreasonable Effectiveness of Mathematics in the Natural Science

"물리 법칙을 정식화하는 데 있어 수학이라는 언어가 이렇게까지
절묘하게 들어맞는 것은 [illegible] 이해하지도 못하고,
감히 받을 자격조차 없는 [illegible] 선물이다."

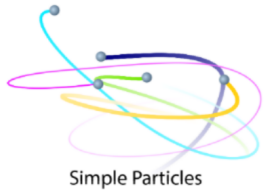
"AI가 데이터만으로 과학 법칙을 스스로 찾아낸다면?"

— Eugene Wigner

"The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve."

The Unreasonable Effectiveness of Mathematics in the Natural Science

Motivation



$a=\{GMm/r^2\}/m$

$a=F/m$

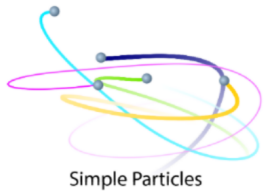
$F=GMm/r^2$

1. Deep Learning
2. Symbolic Regression

복잡한 고차원 데이터에서 '해석 가능한 법칙(수식)'을 어떻게 추출할 것인가?

The Unreasonable Effectiveness of Mathematics in the Natural Science

Motivation



$a=\{GMm/r^2\}/m$

$a=F/m$

$F=GMm/r^2$

복잡한 고차원 데이터에서 '해석 가능한 법칙(수식)'을 어떻게 추출할 것인가?

1. Deep Learning*

- **복잡한 문제**를 잘 맞추고, **계산**이 비교적 가벼움
- Blackbox
  → **interpretability** ⇊
  → **Generalization X**

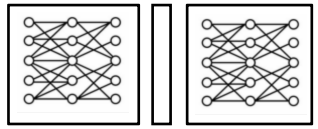
2. Symbolic Regression

- 함수 형태 지정 없이 모델이 어떤 **수학적 형태**인지 찾아냄
  → **interpretability** ⇈
- (유전) 알고리즘으로 여러 함수의 **조합 탐색**
  → **scale exponentially** with # of input variables & operator

*context: Even if we impose strong inductive biases on the models to improve generalization, the learned parts of networks typically are **linear piece-wise approximations** which extrapolate linearly (if using ReLU)

Goal of the Study

Goal of the Study



Inductive biased DL model

\+ Symbolic Regression

A framework that combines the power of **deep learning** with **symbolic regression**

복잡한 고차원 데이터에서 '해석 가능한 법칙(수식)'을 어떻게 추출할 것인가?

1. Deep Learning*

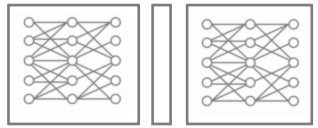
- **복잡한 문제**를 잘 맞추고, **계산**이 비교적 가벼움
- Blackbox
  - → interpretability ⇊
  - → Generalization X

2. Symbolic Regression

- 함수 형태 지정 없이 모델이 어떤 **수학적 형태**인지 찾아냄
  - → interpretability ⇈
- (유전) 알고리즘으로 여러 함수의 **조합 탐색**
  - → **scale exponentially** with # of input variables & operator

Framework Idea

Goal of the Study



Inductive biased DL model

\+ Symbolic Regression

A framework that combines the power of **deep learning** with **symbolic regression**

복잡한 고차원 데이터에서 '해석 가능한 법칙(수식)'을 어떻게 추출할 것인가?

=

Suggested Framework

Step 1. Deep Learning

- Apply strong **inductive biases to DL model (here, GNN)**
  → encourage the learning of **sparse latent representations**
  → 각 NN이 저차원 문제를 해결하게 됨

Step 2. Symbolic Regression

- learned model의 internal parts (NN)에 개별적으로 SR 적용
- Extracts explicit physical equations
  → **interpretability** ↑↑
- 추가적인 여러 inductive bias로 탐색 공간을 대폭 축소

Frame work 제안 → 3 Case studies

Framework Idea

Goal of the Study

Inductive biase

A framework t
deep learning

복잡한 고차원 데이터에서 '해석 가능한 법칙(수식)'을
어떻게 추출할 것인가?

- **Recover the injected closed form physical laws for Newtonian and Hamiltonian examples**
- **Derive a new interpretable closed-form analytical expression that can be useful in astrophysic**

Suggested Framework

Step 1. Deep Learning

)L model (here, GNN)
se latent representations

egression

\)에 개별적으로 SR 적용
IS

→ interpretability ↑↑

- 추가적인 여러 inductive bias로 탐색 공간을 대폭 축소

Frame work 제안 → 3 Case studies

Framework Detail

Step 1. Inductive Bias 설계

- 데이터 특성에 맞는 inductive bias를 제공하는 '분리 가능한 내부 구조'를 가진 딥러닝 모델 설계

Step 2. End-to-End 학습

- **L1 또는 KL 정규화**로 고차원 데이터를 해석 가능한 저차원 희소 표현(**Sparse Representation**) 으로 학습

Step 3. Symbolic Regression

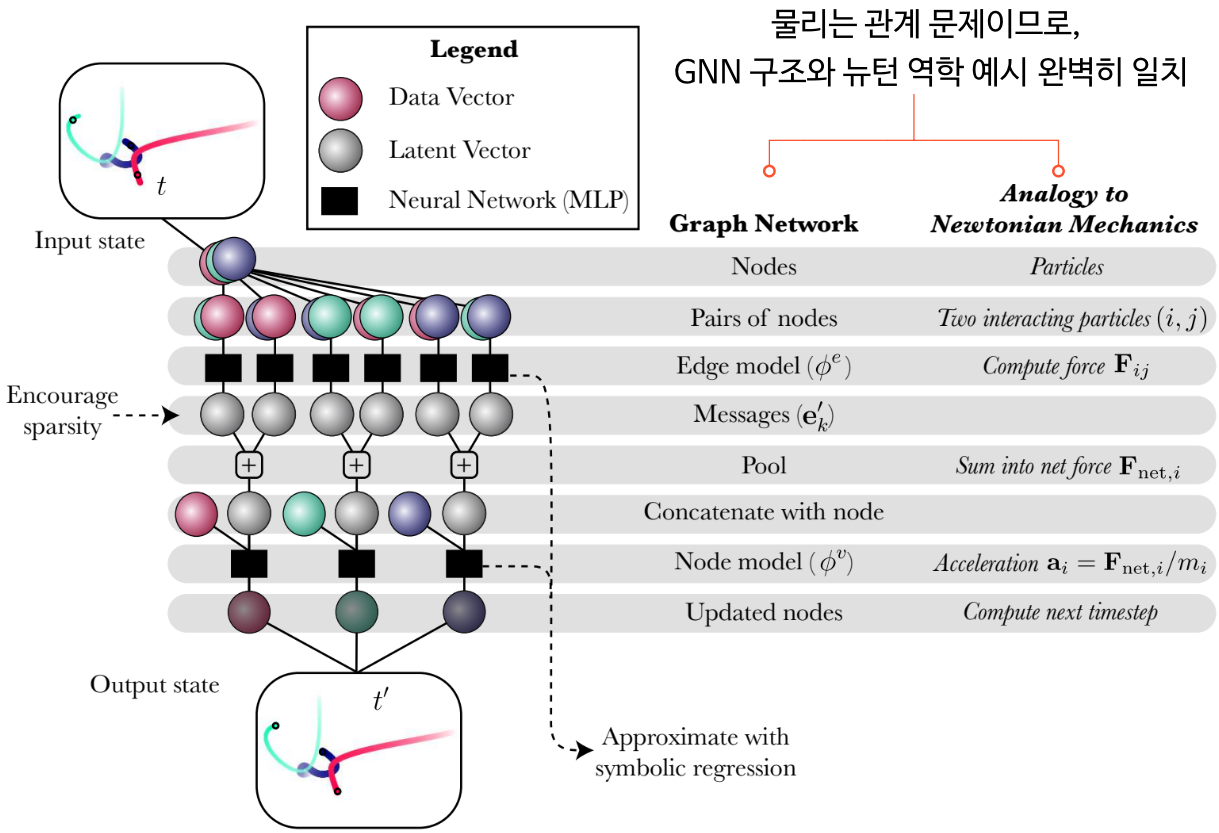
- 학습된 모델에서 내부 구조 (function) 분리
  → 각 NN마다 symbolic regression

Step 4. Mathematical model

- deep model의 해당 함수들을 symbolic 표현으로 대체하여 수리모델을 완성시킴

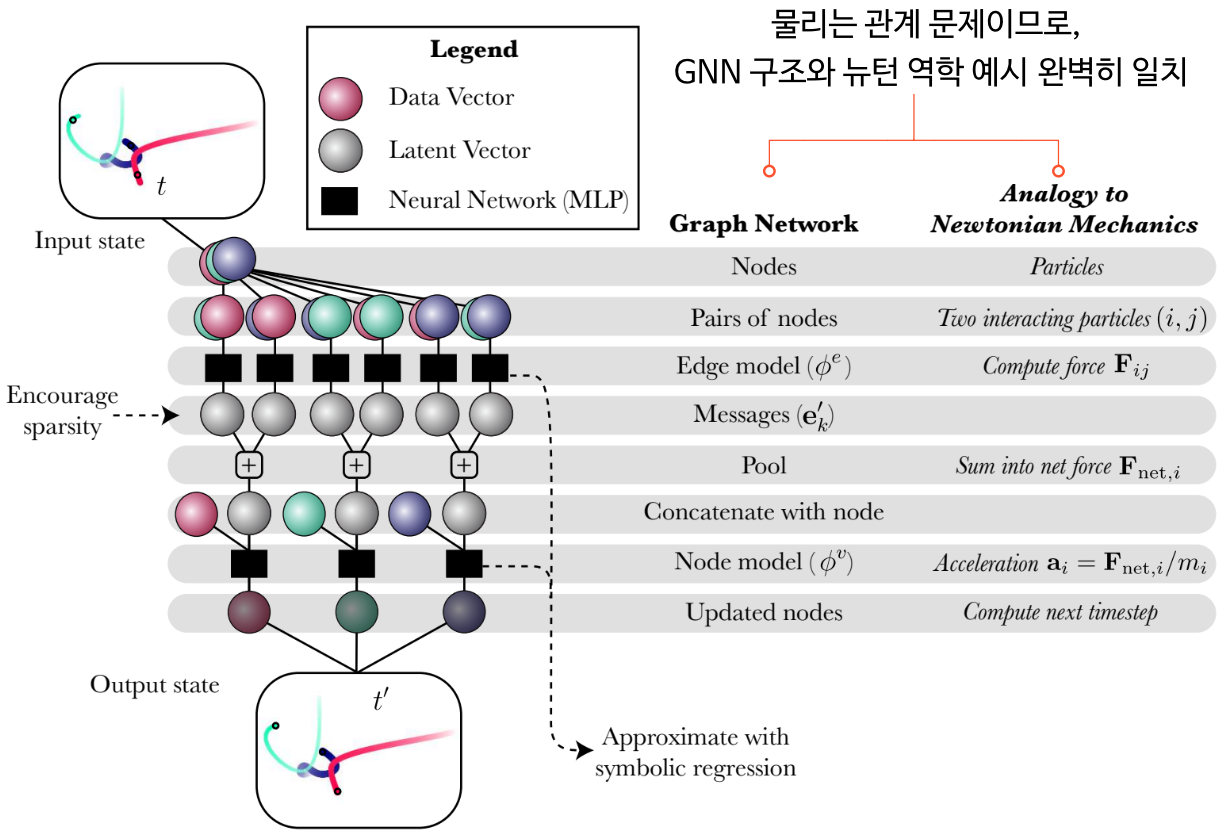
Step 1. Inductive Bias 설계

데이터 특성에 맞는 inductive bias가 주입된 '**분리 가능한 내부 구조**'를 가진 딥러닝 모델 설계



Step 1. Inductive Bias 설계

데이터 특성에 맞는 inductive bias가 주입된 '**분리 가능한 내부 구조**'를 가진 딥러닝 모델 설계



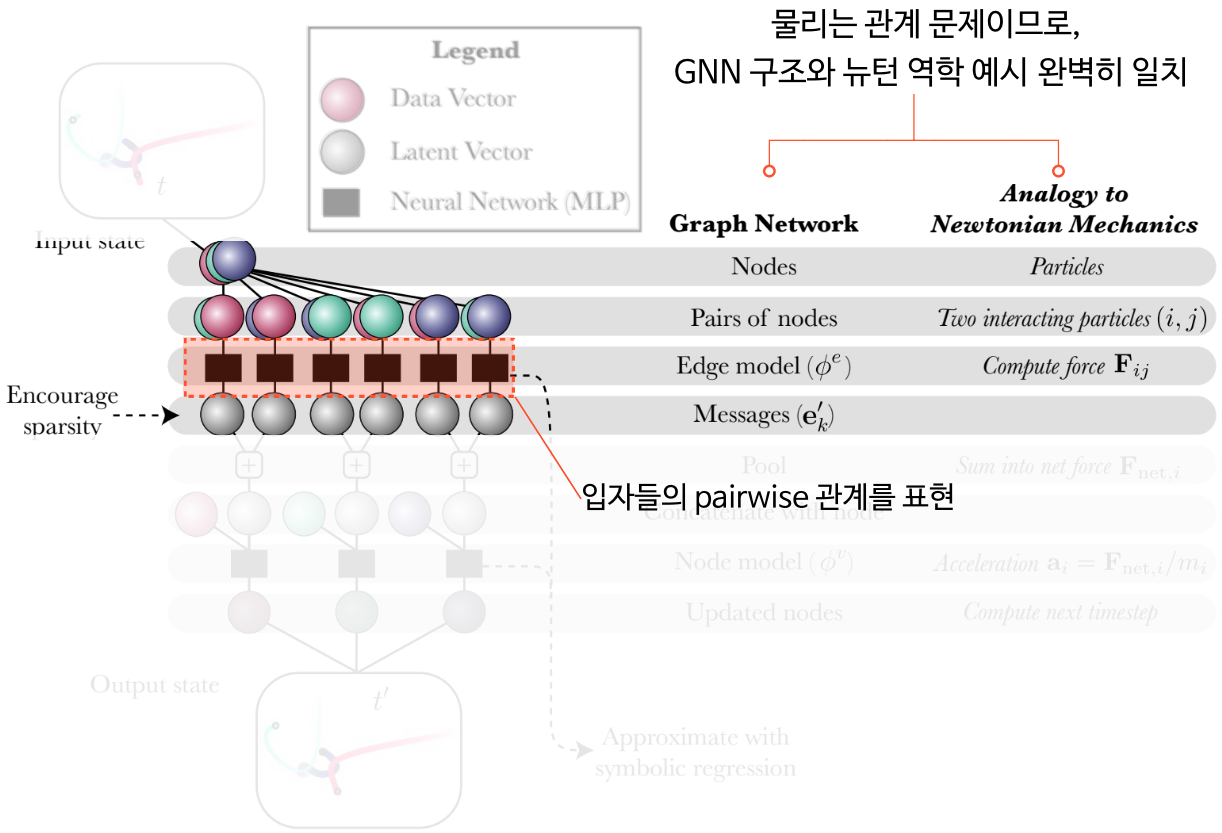
GNN 세 가지 핵심 연산

- **Edge model** (ϕ_e):
  두 노드 간의 관계를 계산하여 메시지를 생성
- **Pool** (Message Aggregation):
  한 노드로 들어오는 모든 메시지를 합침
- **Node model** (ϕ_v):
  합쳐진 정보와 기존 노드 상태를 결합해 노드를 업데이트

$$e_{ij} = \phi_e(v_i, v_j)$$
$$\bar{e}_i = \sum_j e_{ij}$$
$$v'_i = \phi_v(v_i, \bar{e}_i)$$

Step 1. Inductive Bias 설계

데이터 특성에 맞는 inductive bias가 주입된 '분리 가능한 내부 구조'를 가진 딥러닝 모델 설계



물리는 관계 문제이므로,
GNN 구조와 뉴턴 역학 예시 완벽히 일치

입자들의 pairwise 관계를 표현

GNN 세 가지 핵심 연산

- **Edge model** (ϕ_e):
  두 노드 간의 관계를 계산하여 메시지를 생성
- **Pool** (Message Aggregation):
  한 노드로 들어오는 모든 메시지를 합침
- **Node model** (ϕ_v):
  합쳐진 정보와 기존 노드 상태를 결합해 노드를 업데이트

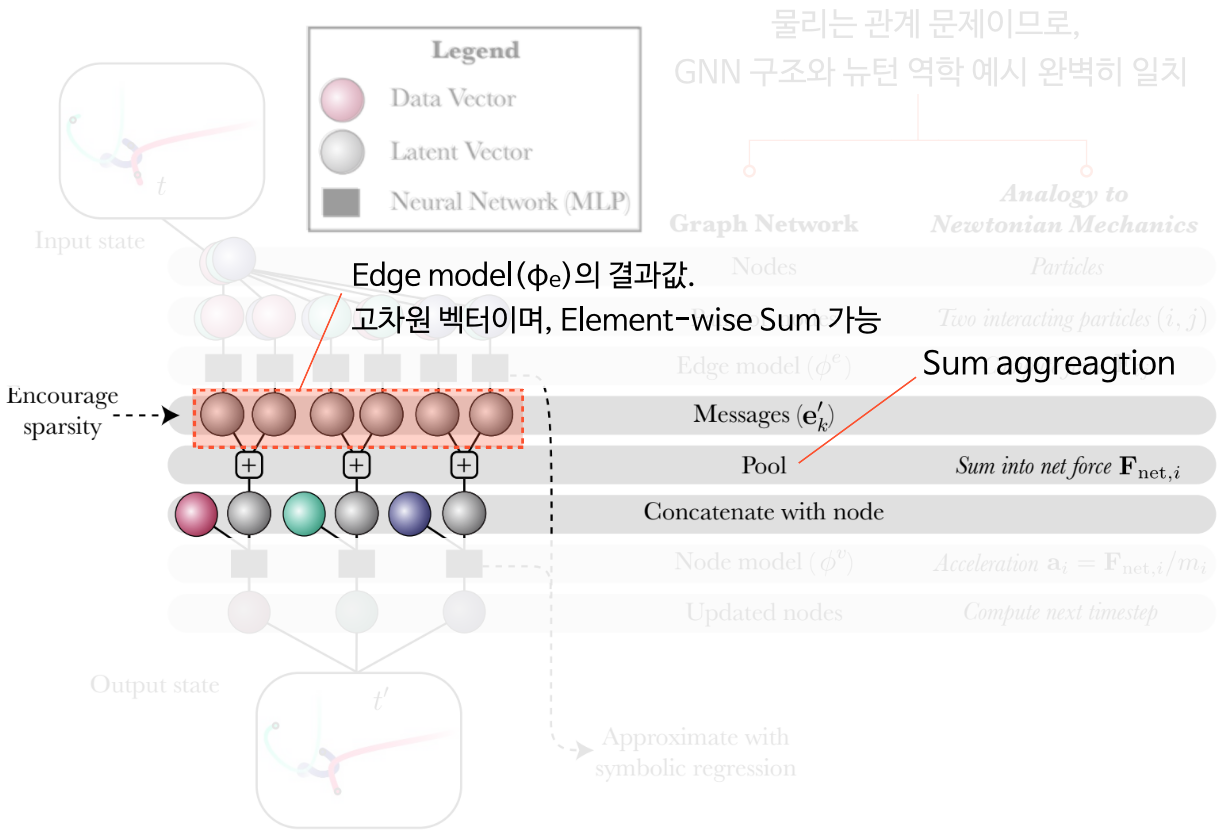
$$e_{ij} = \phi_e(v_i, v_j)$$

$$\bar{e}_i = \sum_j e_{ij}$$

$$v'_i = \phi_v(v_i, \bar{e}_i)$$

Step 1. Inductive Bias 설계

데이터 특성에 맞는 inductive bias가 주입된 '분리 가능한 내부 구조'를 가진 딥러닝 모델 설계



물리는 관계 문제이므로,
GNN 구조와 뉴턴 역학 예시 완벽히 일치

Graph Network | Analogy to Newtonian Mechanics
Nodes | Particles
Two interacting particles (i, j)
Edge model (ϕ^e)
Node model (ϕ^v)
Acceleration $\mathbf{a}_i = \mathbf{F}_{\text{net},i}/m_i$
Updated nodes | Compute next timestep

GNN 세 가지 핵심 연산

- Edge model (ϕe):
  두 노드 간의 관계를 계산하여 메시지를 생성
- **Pool** (Message Aggregation):
  한 노드로 들어오는 모든 메시지를 합침
- Node model (ϕv):
  합쳐진 정보와 기존 노드 상태를 결합해 노드를 업데이트

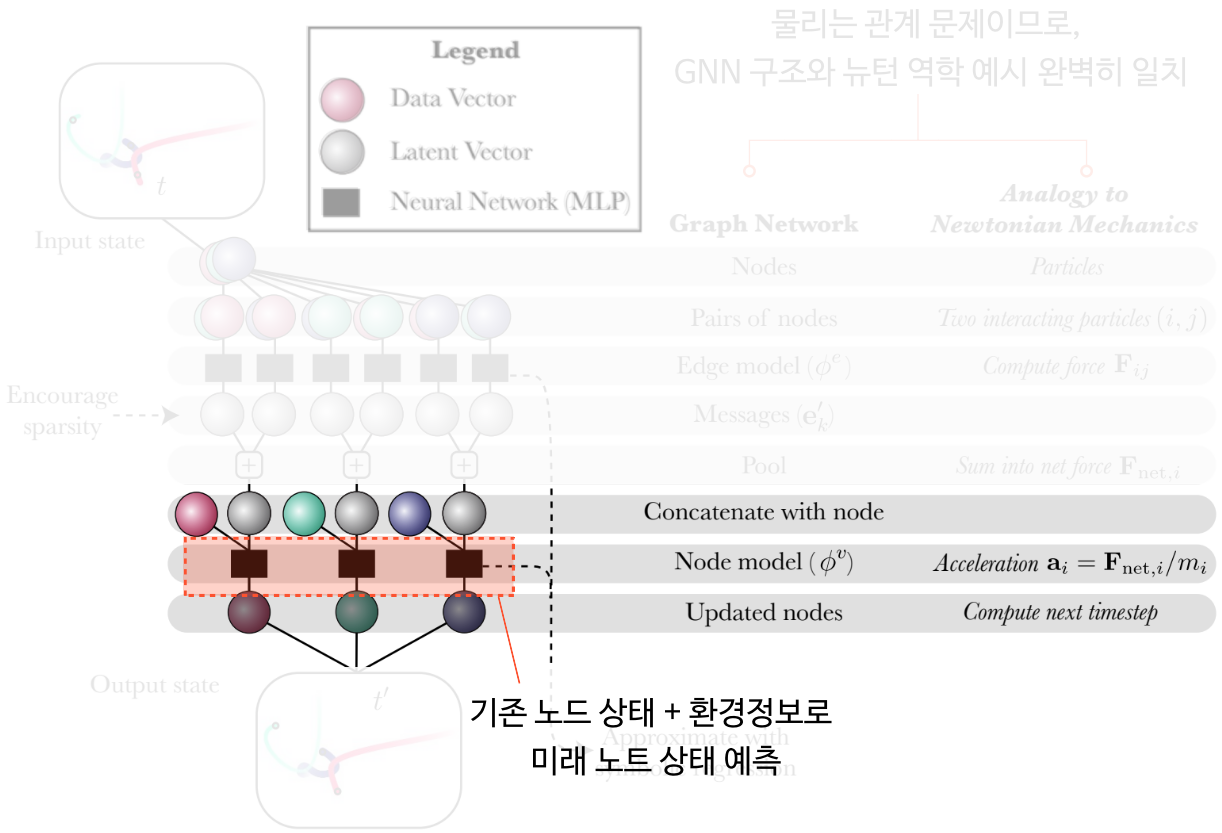
$$e_{ij} = \phi_e(v_i, v_j)$$

$$\bar{e}_i = \sum_j e_{ij}$$

$$v'_i = \phi_v(v_i, \bar{e}_i)$$

Step 1. Inductive Bias 설계

데이터 특성에 맞는 inductive bias가 주입된 '분리 가능한 내부 구조'를 가진 딥러닝 모델 설계



물리는 관계 문제이므로,
GNN 구조와 뉴턴 역학 예시 완벽히 일치

GNN 세 가지 핵심 연산

- Edge model (ϕ_e):
  두 노드 간의 관계를 계산하여 메시지를 생성
- Pool (Message Aggregation):
  한 노드로 들어오는 모든 메시지를 합침
- **Node model (ϕ_v):**
  합쳐진 정보와 기존 노드 상태를 결합해 노드를 업데이트

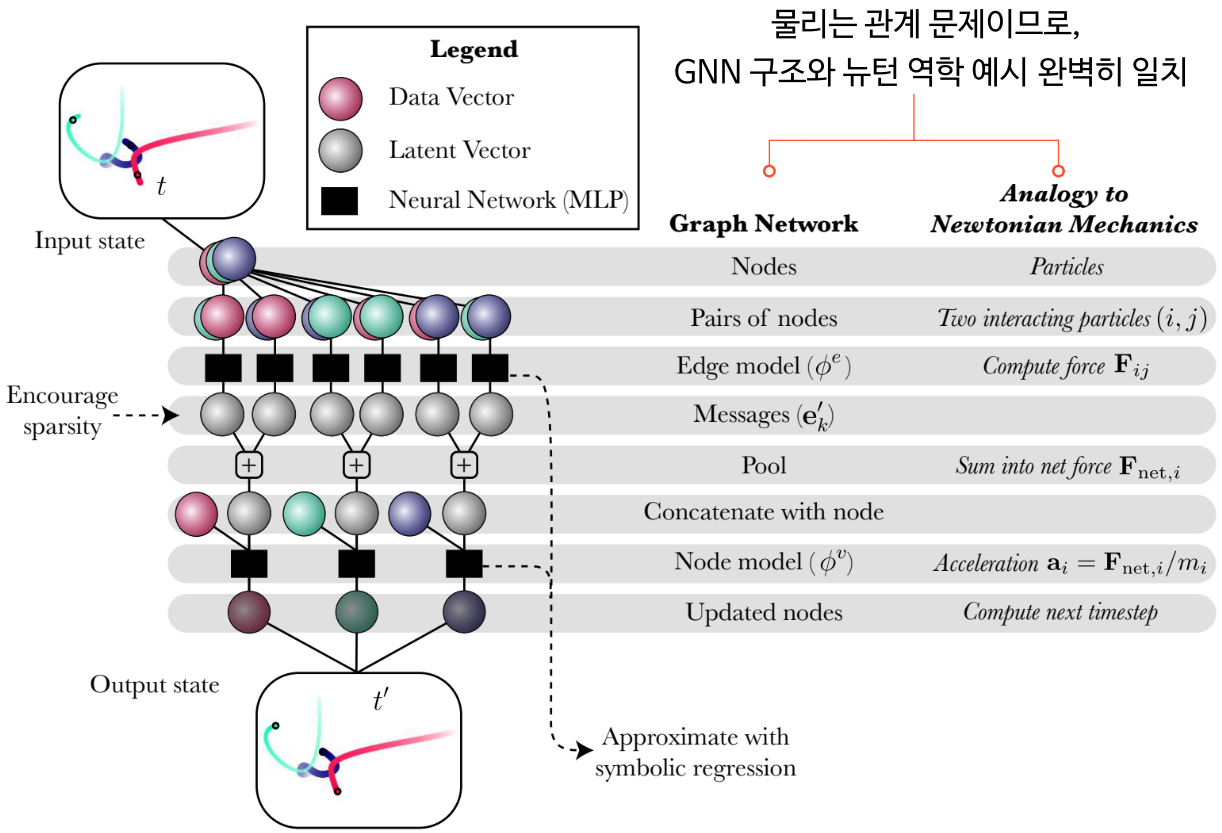
$$e_{ij} = \phi_e(v_i, v_j)$$

$$\bar{e}_i = \sum_j e_{ij}$$

$$v'_i = \phi_v(v_i, \bar{e}_i)$$

Step 1. Inductive Bias 설계

데이터 특성에 맞는 inductive bias가 주입된 '**분리 가능한 내부 구조**'를 가진 딥러닝 모델 설계



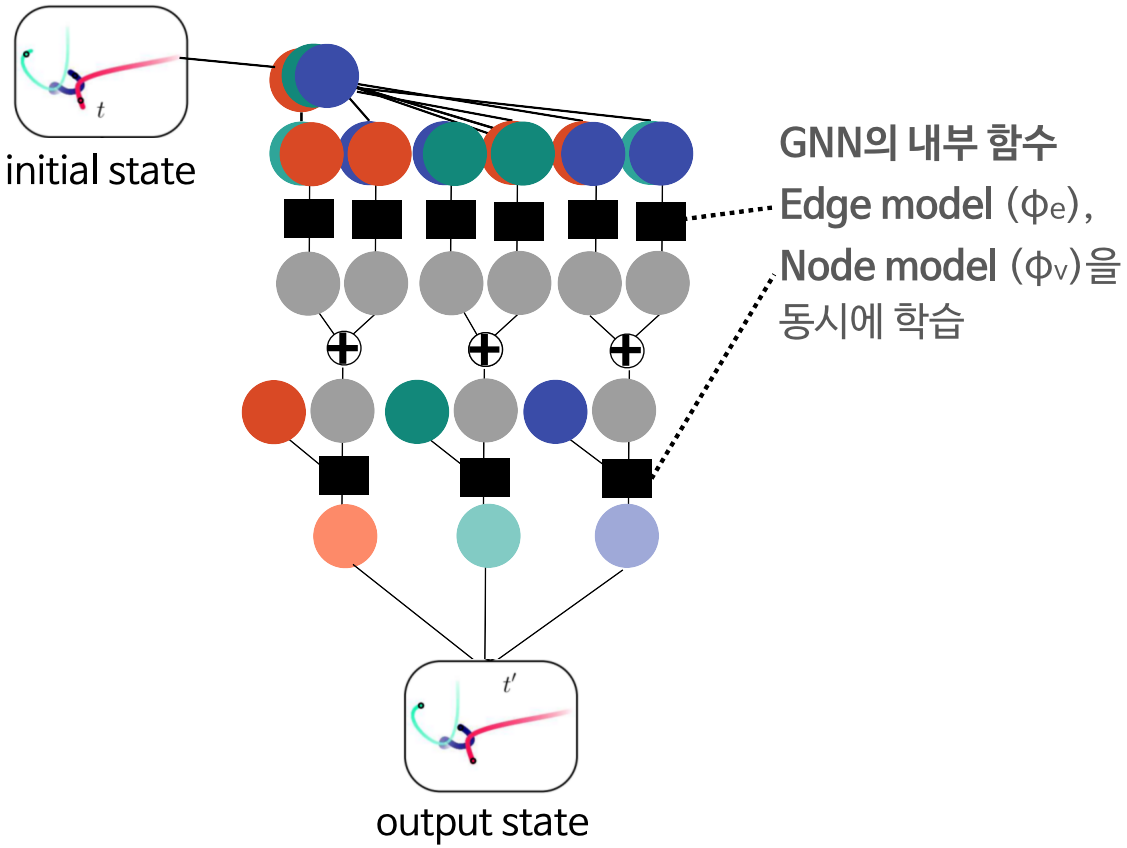
GNN 세 가지 핵심 연산

- **Edge model** (ϕ_e):
  두 노드 간의 관계를 계산하여 메시지를 생성
- **Pool** (Message Aggregation):
  한 노드로 들어오는 모든 메시지를 합침
- **Node model** (ϕ_v):
  합쳐진 정보와 기존 노드 상태를 결합해 노드를 업데이트

$$e_{ij} = \phi_e(v_i, v_j)$$
$$\bar{e}_i = \sum_j e_{ij}$$
$$v'_i = \phi_v(v_i, \bar{e}_i)$$

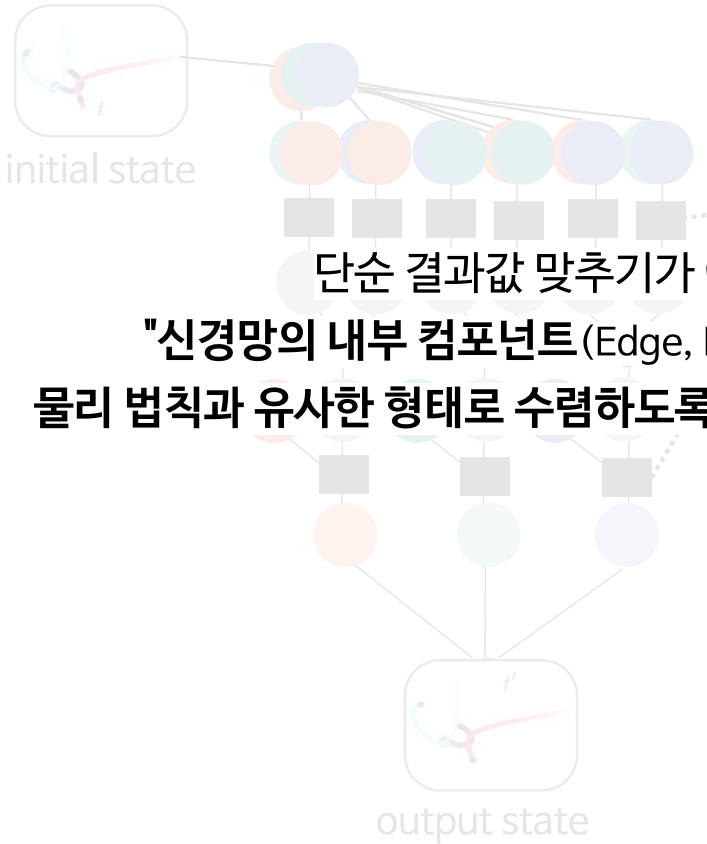
Step 2. End-to-End 학습

L1 또는 KL 정규화로 고차원 데이터를 해석 가능한 저차원 희소 표현(Sparse Representation) 으로 학습



Step 2. End-to-End 학습

L1 또는 KL 정규화로 고차원 데이터를 해석 가능한 저차원 희소 표현(Sparse Representation) 으로 학습

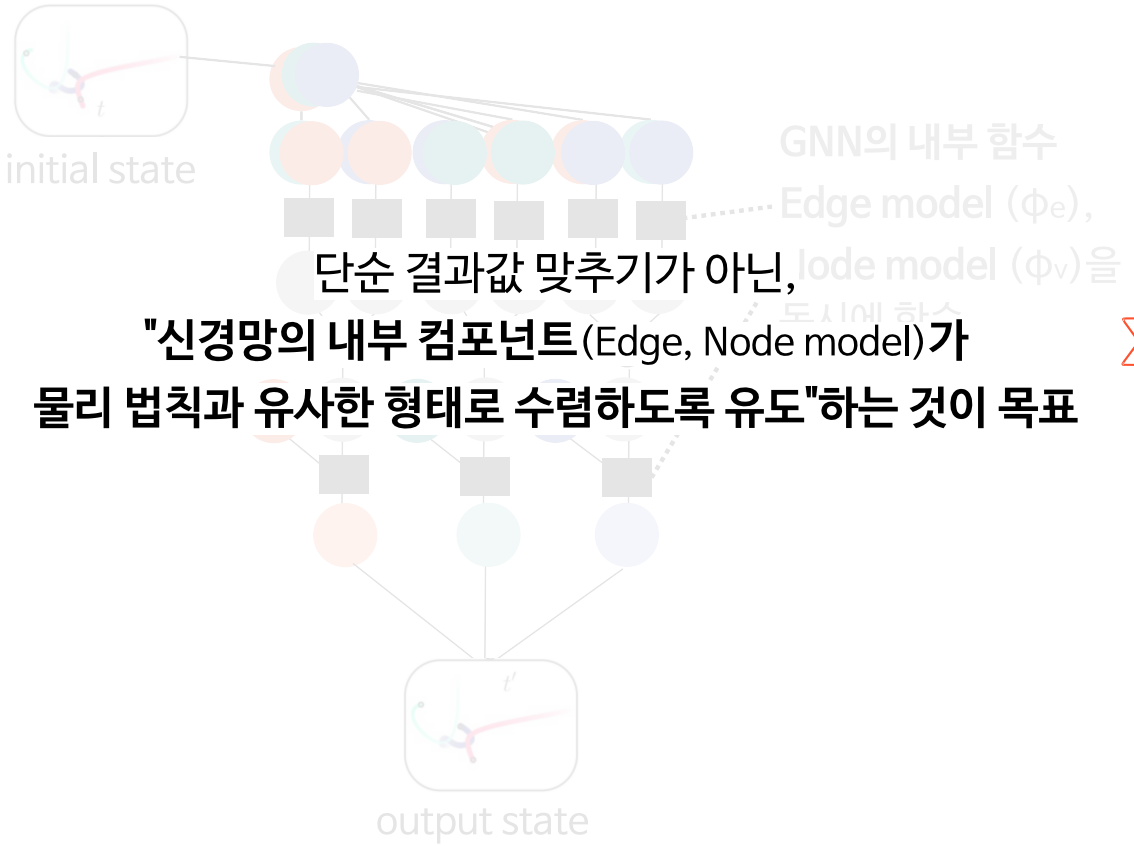


GNN의 내부 함수
Edge model (ϕ_e),
lode model (ϕ_v)을

단순 결과값 맞추기가 아닌,
"신경망의 내부 컴포넌트(Edge, Node model)가
물리 법칙과 유사한 형태로 수렴하도록 유도"하는 것이 목표

Step 2. End-to-End 학습

L1 또는 KL 정규화로 고차원 데이터를 해석 가능한 저차원 희소 표현(Sparse Representation) 으로 학습



단순 결과값 맞추기가 아닌,
"신경망의 내부 컴포넌트(Edge, Node model)가
물리 법칙과 유사한 형태로 수렴하도록 유도"하는 것이 목표

Total Loss Function

$$\mathcal{L} = \mathcal{L}_v + \alpha_1 \mathcal{L}_e + \alpha_2 \mathcal{L}_n$$

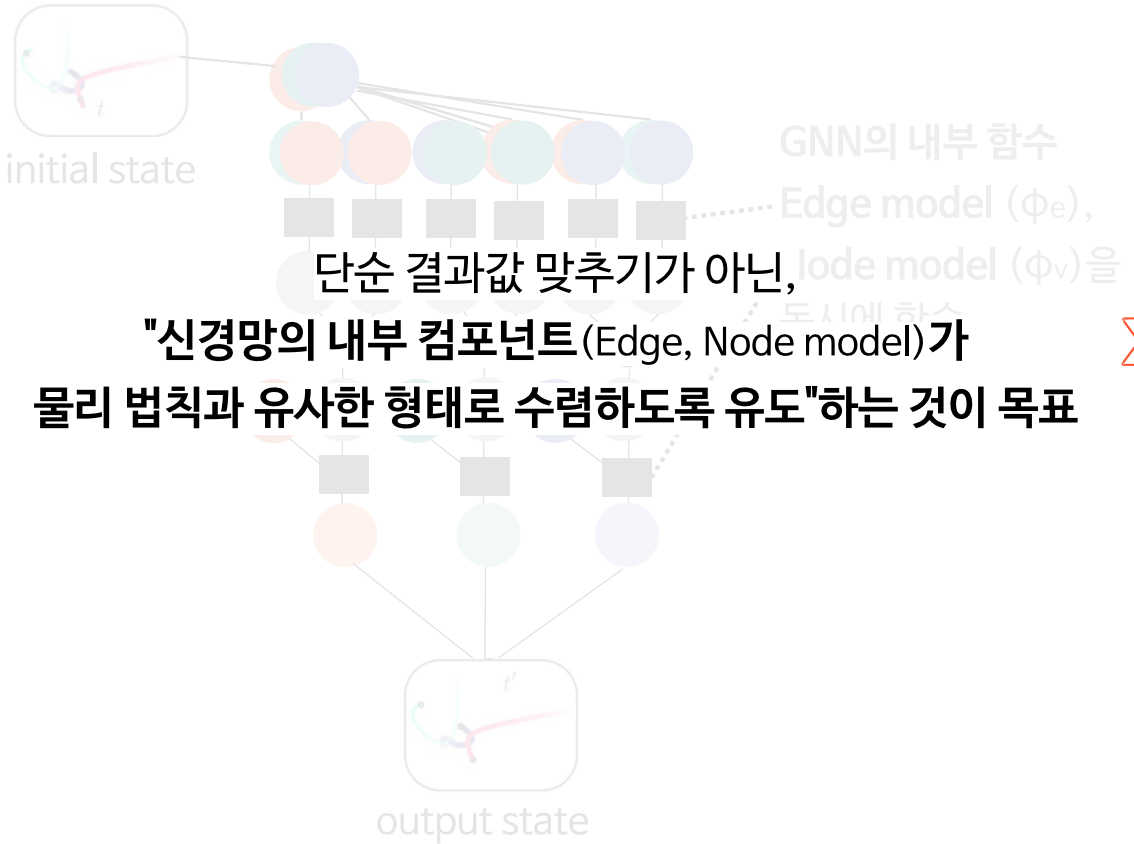
$\mathcal{L}_v$: Prediction Loss (MAE)

$\mathcal{L}_e$: Message Regularization (L1, KL, Bottleneck...)

$\mathcal{L}_n$: Weight Decay
- 모델의 복잡도를 제어하는 일반적인 가중치 규제

Step 2. End-to-End 학습

L1 또는 KL 정규화로 고차원 데이터를 해석 가능한 저차원 희소 표현(Sparse Representation) 으로 학습



단순 결과값 맞추기가 아닌,
"신경망의 내부 컴포넌트(Edge, Node model)**가**
물리 법칙과 유사한 형태로 수렴하도록 유도"하는 것이 목표

Sparsify: How & Why

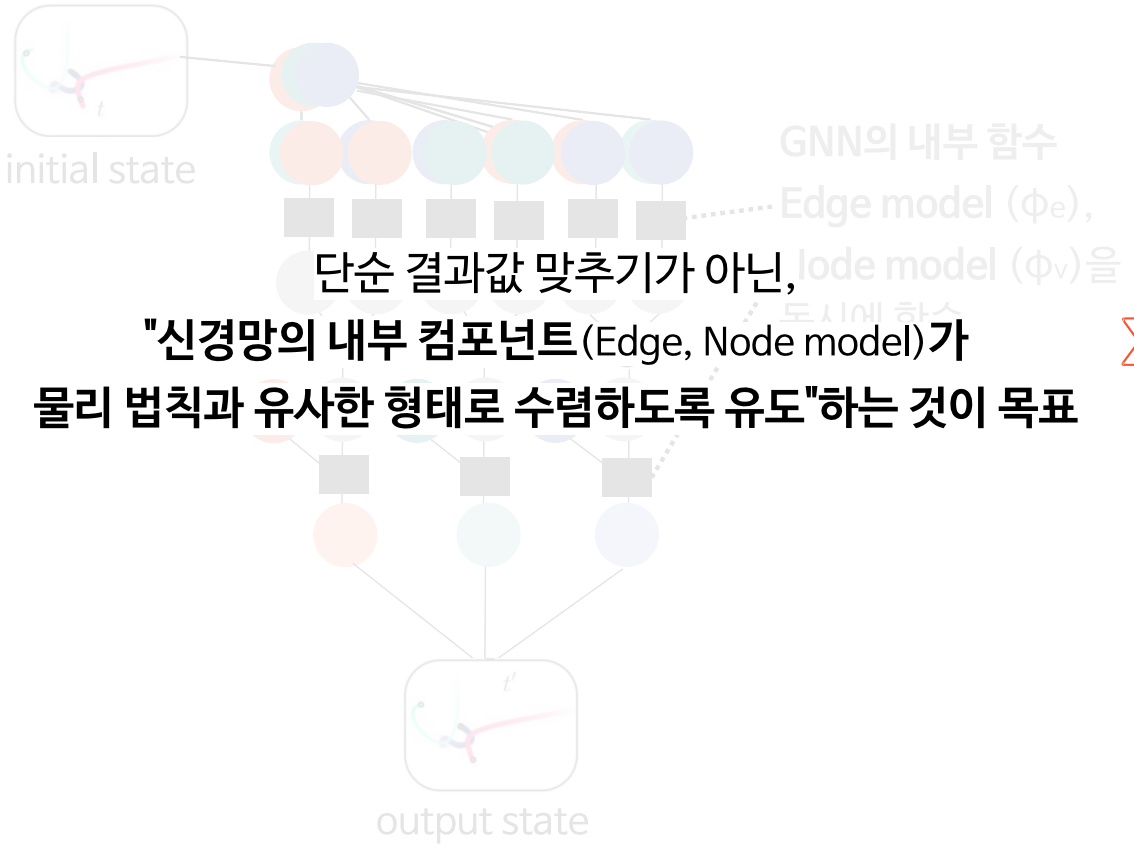
$\mathcal{L}_e$: Message Regularization (L1, KL, Bottleneck...)

$\mathcal{L}_n$: Weight Decay
- 모델의 복잡도를 제어하는 일반적인 가중치 규제

① **Occam's Razor:** science always prefers the simpler model or representation of two which give similar accuracy

Step 2. End-to-End 학습

L1 또는 KL 정규화로 고차원 데이터를 해석 가능한 저차원 희소 표현(Sparse Representation) 으로 학습



단순 결과값 맞추기가 아닌,
"신경망의 내부 컴포넌트(Edge, Node model)**가**
물리 법칙과 유사한 형태로 수렴하도록 유도"하는 것이 목표

Sparsify: How & Why

$\mathcal{L}_e$: Message Regularization (L1, KL, Bottleneck...)

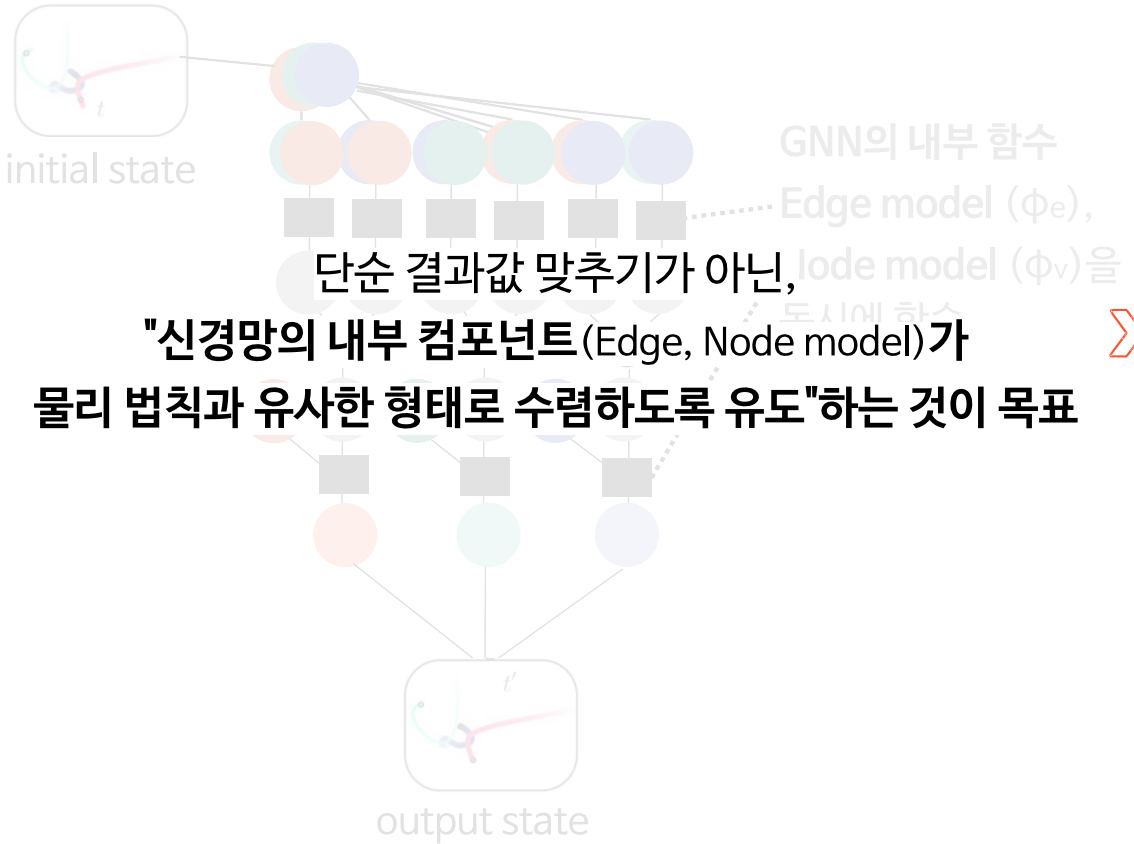
$\mathcal{L}_n$: Weight Decay
- 모델의 복잡도를 제어하는 일반적인 가중치 규제

① Occam's Razor

② **Representations의 수학적 동등성:**
If a system is perfectly described by messages **in a compact space**,
Learned GNN messages become **mathematical rotations of the true vectors.**

Step 2. End-to-End 학습

L1 또는 KL 정규화로 고차원 데이터를 해석 가능한 저차원 희소 표현(Sparse Representation) 으로 학습



단순 결과값 맞추기가 아닌,
"신경망의 내부 컴포넌트(Edge, Node model)가
물리 법칙과 유사한 형태로 수렴하도록 유도"하는 것이 목표

Sparsify: How & Why

$\mathcal{L}_e$: Message Regularization (L1, KL, Bottleneck...)

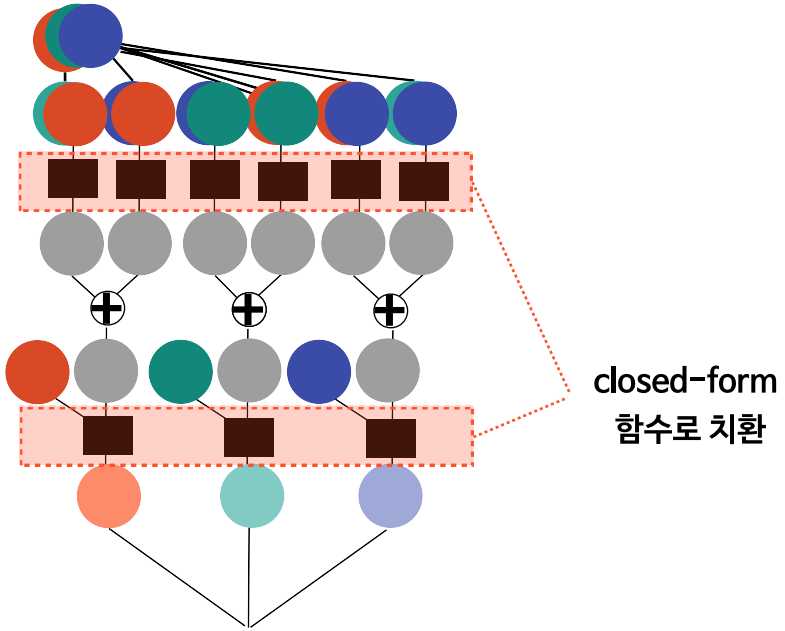
$\mathcal{L}_n$: Weight Decay
- 모델의 복잡도를 제어하는 일반적인 가중치 규제

100차원 메시지 벡터를 사용하더라도, 학습 후에는
실제 물리 차원 수준의 소수의 성분만이 정보를 담게 됨.*

Step 3. Symbolic Regression

학습된 모델에서 내부 구조 (function) 분리 → 각 NN마다 symbolic regression

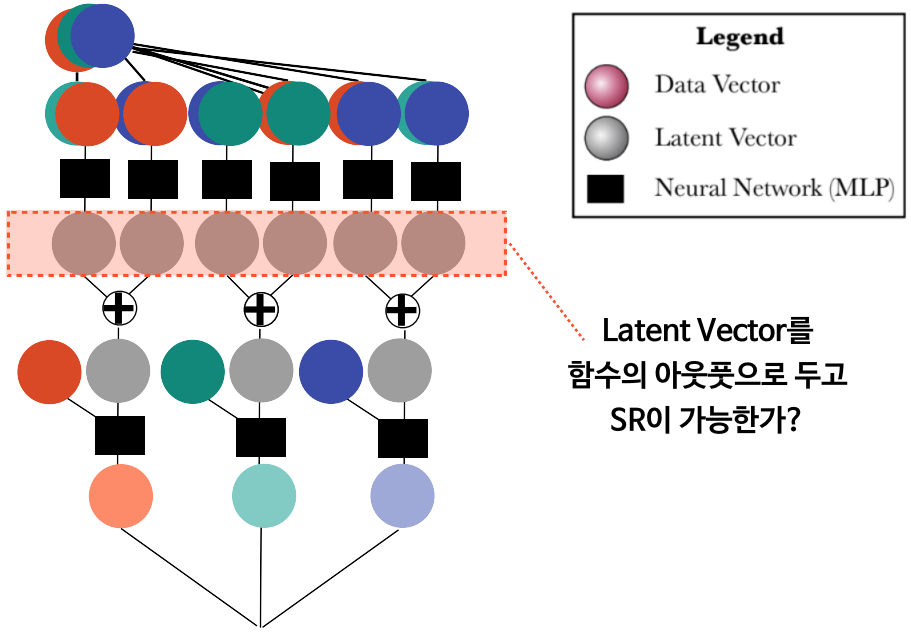
- Fit compact **closed-form analytical expressions** to Φ_e, Φ_v (independently) by eureqa package
  - Eureqa는 유전 알고리즘으로 대수적 표현 조합을 탐색
  - 사용 operators: +, −, ×, /, 〉, 〈, ^, exp, log, IF(·, ·, ·)
  - SR로 대략적 수식 형태 찾으면 계수들은 데이터 피팅으로 찾음



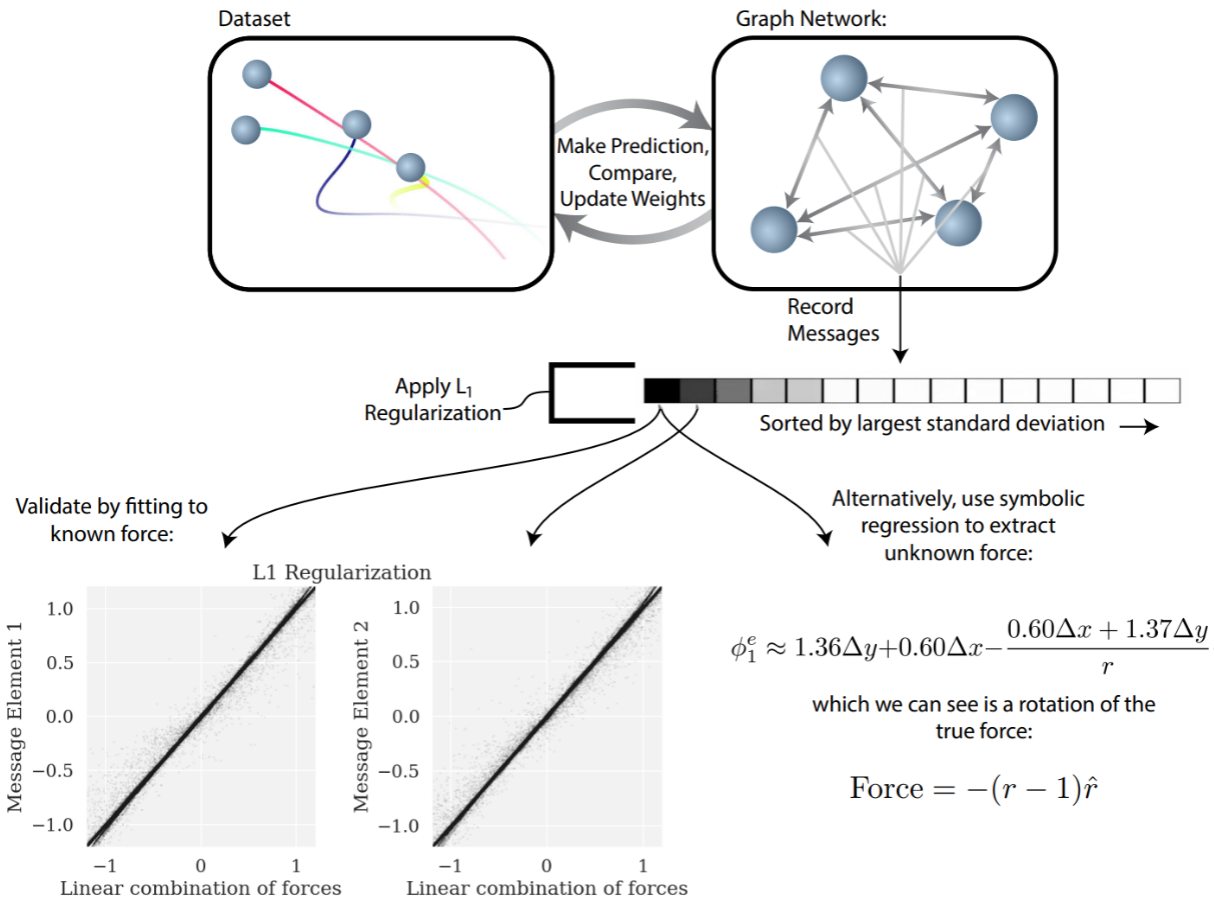
Step 3. Symbolic Regression

학습된 모델에서 내부 구조 (function) 분리 → 각 NN마다 symbolic regression

- Fit compact **closed-form analytical expressions** to Φ_e, Φ_v (independently) by eureqa package
  - Eureqa는 유전 알고리즘으로 대수적 표현 조합을 탐색
  - 사용 operators: +, −, ×, /, 〉, 〈, ^, exp, log, IF(·, ·, ·)
  - SR로 대략적 수식 형태 찾으면 계수들은 데이터 피팅으로 찾음



Case studies – Newtonian dynamics



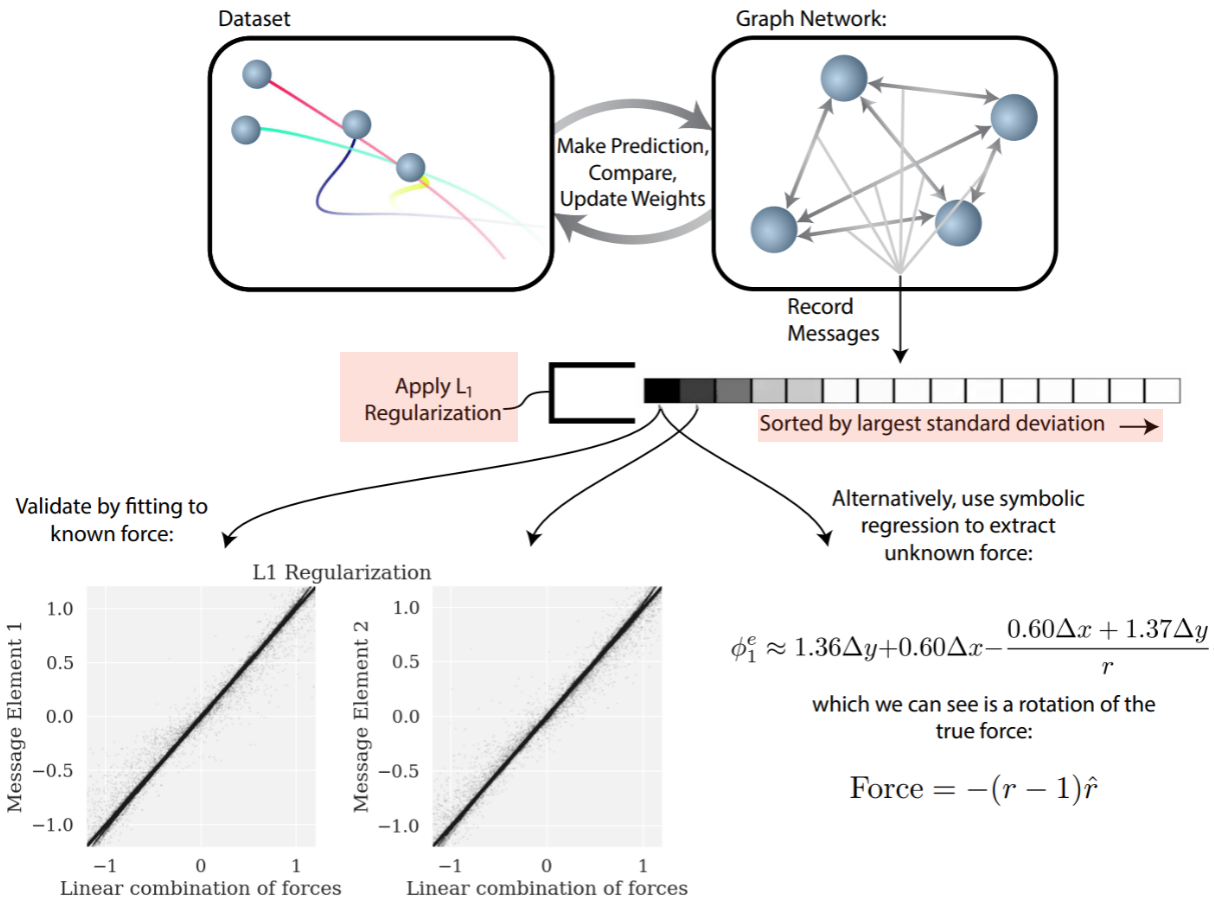
$$\phi_1^e \approx 1.36\Delta y + 0.60\Delta x - \frac{0.60\Delta x + 1.37\Delta y}{r} - 0.0025$$

which we can see is a rotation of the true force:

$$\text{Force} = -(r-1)\hat{r}$$

학습 아키텍처는 acceleration을 맞추도록 함.
그러나 진짜 목표는 ϕ_e안에 숨은 '힘 공식'을 찾는 것

Case studies – Newtonian dynamics



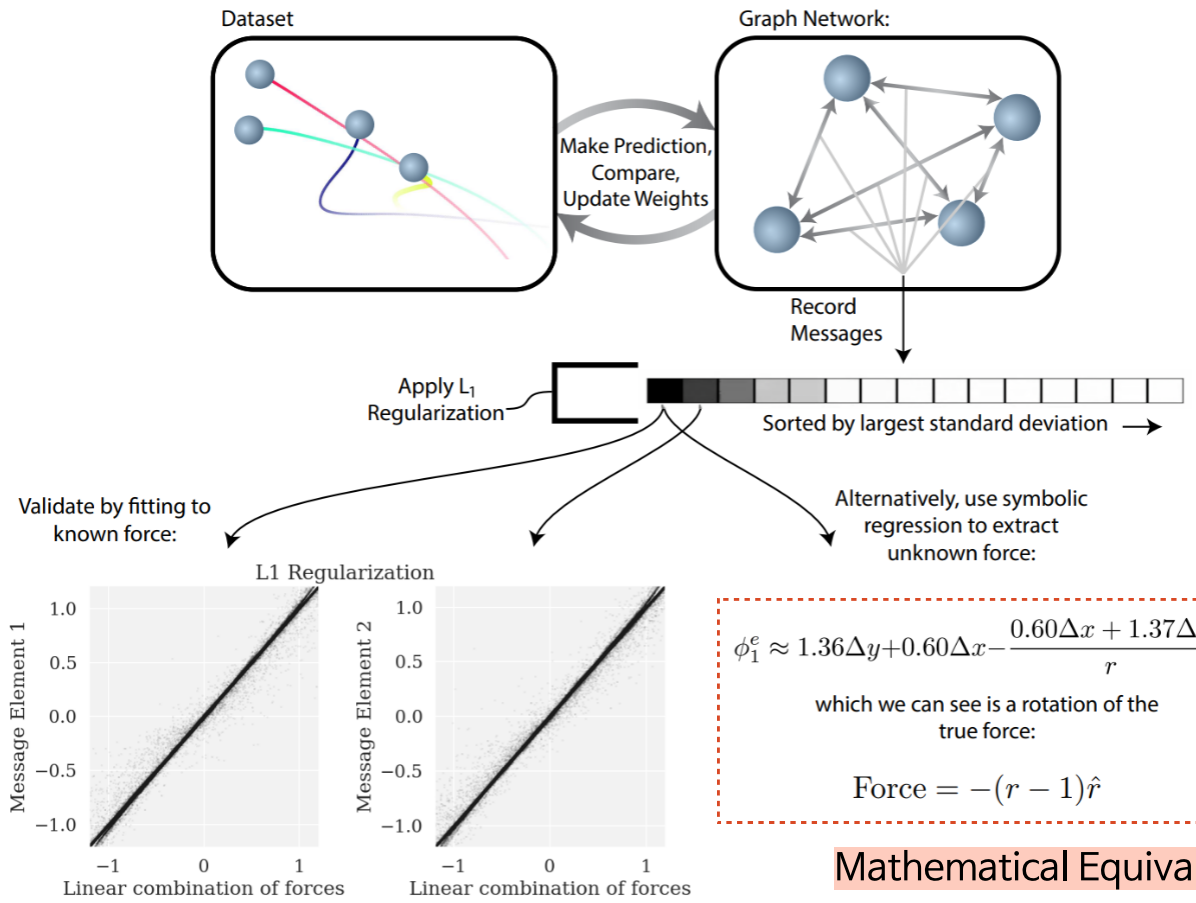
$$\phi_1^e \approx 1.36\Delta y + 0.60\Delta x - \frac{0.60\Delta x + 1.37\Delta y}{r} - 0.0025$$

which we can see is a rotation of the true force:

$$\text{Force} = -(r-1)\hat{r}$$

학습 아키텍처는 acceleration을 맞추도록 함.
그러나 진짜 목표는 ϕ_e안에 숨은 '힘 공식'을 찾는 것

Case studies – Newtonian dynamics



$$\phi_1^e \approx 1.36\Delta y + 0.60\Delta x - \frac{0.60\Delta x + 1.37\Delta y}{r} - 0.0025$$

which we can see is a rotation of the true force:

$$\text{Force} = -(r-1)\hat{r}$$

Mathematical Equivalence !

학습 아키텍처는 acceleration을 맞추도록 함.
그러나 진짜 목표는 ϕe안에 숨은 '힘 공식'을 찾는 것

시뮬레이션에 사용된 6가지 물리 법칙

(a) 1/r orbital force: 행성 운동 등에서 나타나는 중심력 기반의 궤도 운동
(b) $1/r^2$ orbital force: 일반적인 만유인력 법칙
(c) Coulomb force
(d) Damped Springs: 용수철의 힘에 속도에 비례하는 저항(damping)이 추가된 모델
(e) Discontinuous forces: 입자 간 거리가 2보다 작아지면 힘이 0이 되는,
즉 특정 조건에서만 작동하는 까다로운 물리 법칙
(f) 전체 입자 간 스프링: 모든 입자가 서로 용수철로 연결된 복잡한 상호작용 시스템

Case studies – Newtonian dynamics

Message Regularization during Model Training

- **Standard:**
  제약 없이 100개의 메시지 성분을 모두 사용하는 일반적인 GNN
- **구조적 제약 (Bottleneck):**
  메시지 차원을 물리적 차원(2D면 2차원, 3D면 3차원)과 똑같이 제한
- **수학적 제약 (L_1):**
  100 차원에 L1 손실 추가 → 대부분 차원 값이 0되고, 1~2개의 차원만 생존.
- **관성적 제약 (KL Divergence):**
  아무 정보가 없는 기본상태의 표준 정규분포를 사전분포로 설정
  → 학습중인 메시지 분포가 이 prior과 다를수록 패널티를 부여

Sim.	Standard	Bottleneck	L_1	KL
Charge-2	0.016	0.947	0.004	0.185
Charge-3	0.013	0.980	0.002	0.425
r^{-1}-2	0.000	1.000	1.000	0.796
r^{-1}-3	0.000	1.000	1.000	0.332
r^{-2}-2	0.004	0.993	0.990	0.770
r^{-2}-3	0.002	0.994	0.977	0.214
Spring-2	0.032	1.000	1.000	0.883
Spring-3	0.036	0.995	1.000	0.214

Sim.	Standard	Bottleneck	L_1	KL	FlatHGN
Charge-2	**49**	50	52	60	55
Charge-3	1.2	0.99	**0.94**	4.2	3.5
Damped-2	**0.30**	0.33	**0.30**	1.5	0.35
Damped-3	0.41	0.45	**0.40**	3.3	0.47
Disc.-2	0.064	0.074	**0.044**	1.8	0.075
Disc.-3	0.20	0.18	**0.13**	4.2	0.14
r^{-1}-2	0.077	0.069	0.079	3.5	**0.05**
r^{-1}-3	0.051	0.050	0.055	3.5	**0.017**
r^{-2}-2	1.6	1.6	**1.2**	9.3	1.3
r^{-2}-3	4.0	3.6	3.4	9.8	**2.5**
Spring-2	0.047	0.046	0.045	1.7	**0.016**
Spring-3	0.11	0.11	0.090	3.8	**0.010**

Case studies – Dark matter halos

Problem Definition

- **배경**: 우주에는 안 보이지만 큰 질량을 가진 암흑 물질(Dark Matter)이 존재.
  이 물질들은 자기들끼리 뭉쳐서 'Halo'라는 덩어리를 형성.
- **목표 변수 (δ)**: 특정 지역에 암흑 물질이 얼마나 '빽빽하게' 뭉쳐 있는지(overdensity)를 예측.
- **입력 변수 (M, r, v)**: 그 지역에 있는 할로의 질량, 위치, 속도 데이터

Why GNN?

우주는 거대한 네트워크 구조임. 할로들은 서로 중력으로 연결되어 영향 받기 때문.

Case studies – Dark matter halos

Problem Definition

- **배경**: 우주에는 안 보이지만 큰 질량을 가진 암흑 물질(Dark Matter)이 존재.
  이 물질들은 자기들끼리 뭉쳐서 'Halo'라는 덩어리를 형성.
- **목표 변수 (δ)**: 특정 지역에 암흑 물질이 얼마나 '빽빽하게' 뭉쳐 있는지(overdensity)를 예측.
- **입력 변수 (M, r, v)**: 그 지역에 있는 할로의 질량, 위치, 속도 데이터

	Test	Formula	Summed Component	$\langle \lvert \delta_i - \hat{\delta}_i \rvert \rangle$
Old	Constant	$\hat{\delta}_i = C_1$	N/A	0.421
	Simple	$\hat{\delta}_i = C_1 + (C_2 + M_i C_3) e_i$	$e_i = \sum_{j \neq i}^{\lvert \mathbf{r}_i - \mathbf{r}_j \rvert < 20} M_j$	0.121
New	Best, without mass	$\hat{\delta}_i = C_1 + \frac{e_i}{C_2 + C_3 e_i \lvert \mathbf{v}_i \rvert}$	$e_i = \sum_{j \neq i} \frac{C_4 + \lvert \mathbf{v}_i - \mathbf{v}_j \rvert}{C_5 + (C_6 \lvert \mathbf{r_i} - \mathbf{r_j} \rvert)^{C_7}}$	0.120
	Best, with mass	$\hat{\delta}_i = C_1 + \frac{e_i}{C_2 + C_3 M_i}$	$e_i = \sum_{j \neq i} \frac{C_4 + M_j}{C_5 + (C_6 \lvert \mathbf{r}_i - \mathbf{r}_j \rvert)^{C_7}}$	0.0882

찾아낸 식의 물리적 원리/의의는 분석해내지 못함

물리학자들의 직관보다
더 뛰어난 예측력을 가진 fomula

거리가 멀어질수록 영향력이 줄어든다는
물리적 상식을 AI가 수식 형태에 그대로 반영

Conclusion: Summary of Contributions

1. 물리적 Inductive Bias를 통한 해석 가능성 확보

- 신경망 내부 메시지가 실제 물리량(힘, 에너지)의 선형 변환과 동등함을 증명.
- 블랙박스 연산을 이해 가능한 물리적 개념으로 연결

2. GNN과 Symbolic Regression의 결합 (Factorization)

- 훈련된 모델의 내부 함수에서 SR을 통해 명시적인 수식을 추출하는기법 제시
- GNN의 명시적인 Substructure → 일반적인 신경망보다 훨씬 세밀한 해석과 공식 추출이 가능

3. 실제 과학적 발견으로의 확장성 입증

- 단순 시뮬레이션을 넘어, Dark Matter 문제를 풀어기존 공식을 능가하는 새로운 수식 발견
- 본 프레임워크가 미지의 과학적 원리를 탐구하는 강력한 도구가 될 수 있음을 실증함.