Trustworthy AI for Human and Science

Yongfeng Zhang, Rutgers University

yongfeng.zhang@rutgers.edu

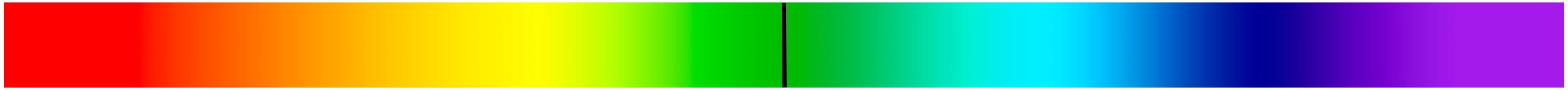
http://www.yongfeng.me

RUTGERS

AI helps in many Research Areas

- A (very rough) spectrum of research discipline system

Human/Arts Nature/Science



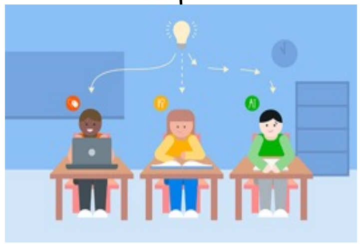
... Art Literature Sociology Education Economics CS/AI Engineering Biology Medical Chemistry Physics ...

AI Painting

WHAT STARTS OUT AS TRIUMPH SOON BECOMES MANIPULATED INTO A CACOPHONY OF FUTILITY, LEAVING ONLY A SENSE OF DREAD AND THE UNLIKELI-HOOD OF A NEW ORDER.

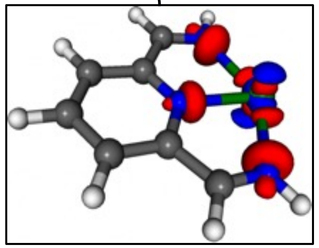
BY TEXT GENERATOR

AI-assisted writing



AI-assisted Learning

AI for Biology, AlphaFold

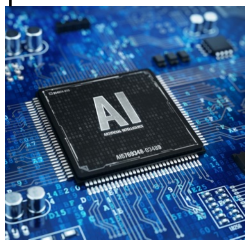


AI for Chemistry

AI and Arts

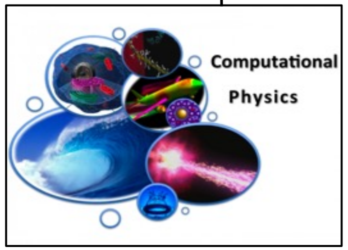
AI for Social Science

Computational Econ



AI for Chip Design

AI for Drug Discovery



AI for Physics

Trustworthy AI

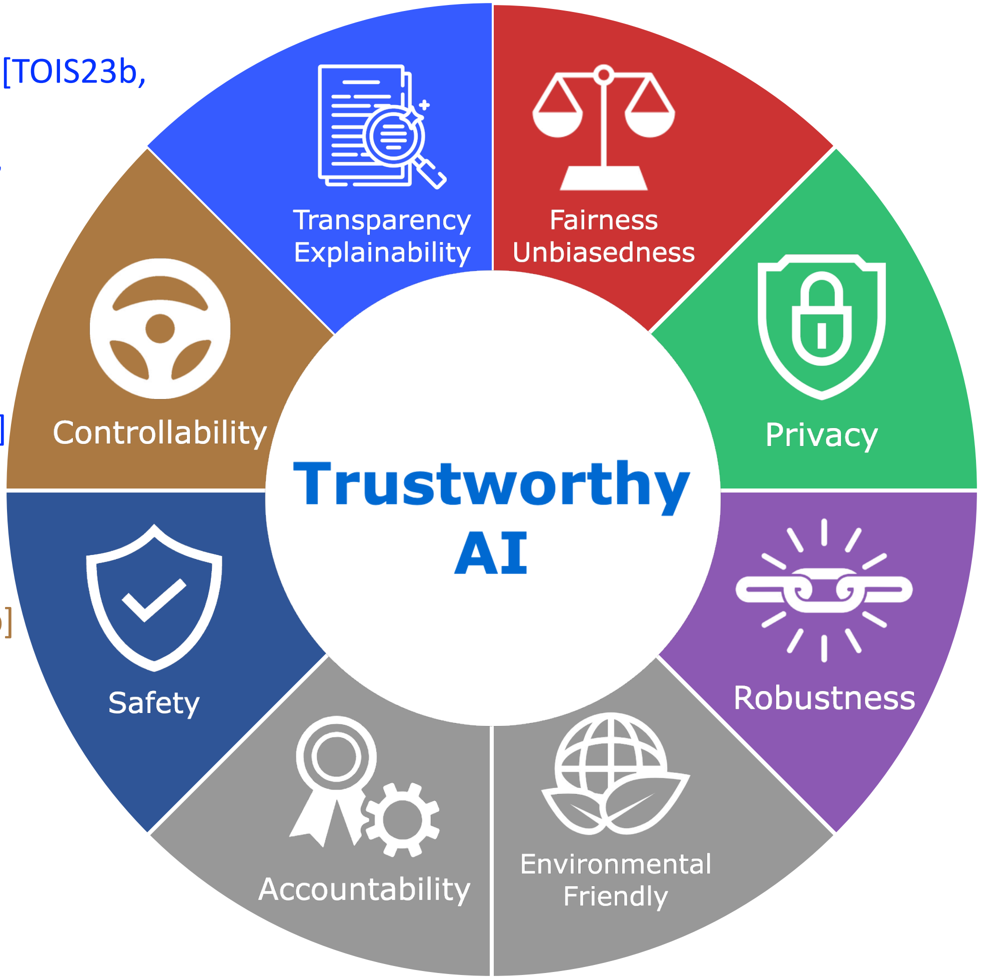
RUTGERS

Our Research Landscape – Methodology

Counterfactual Explanation [SIGIR23, KDD23, ECAI23, WSDM23, WWW22a, CIKM21a]
Natural Language Explanation [CIKM23, EMNLP22b, ACL21, CIKM20c, COLING20, SIGIR14, AAAI19]
Visual Explanation [ICLR23, ACL22, SIGIR19b]
Large Language Model (LLM) based Explanation [TOIS23b, CIKM23, RecSys22a]
Neural-Symbolic Explanation [ICLM22, SIGIR22b, EMNLP22a, WSDM22a, CIKM22b, WWW21a, NAACL21, CIKM20a, CIKM20b]
Knowledge-based Explanation [WWW22b, RecSys21, SIGIR21c, SIGIR19a]
Explainable Model Debugging [TACL23, TOIS23a]
Evaluation of Explanations [WWW22c, SIGIR21d]

Controllable Text Generation [WWW20, ACL21, RecSys22a, TOIS23b, CIKM20c]
Controllable Image Generation [ACL22, SIGIR19b]
User Controllable Recommendation [ECAI23]
User Controllable Fairness [SIGIR21a]
White-box based Controllability [TOIS23a]

Counterfactual Attacking [SIGIR23]
Shilling Attack Detection [IJCAI15]



Counterfactual Fairness [SIGIR21a]
User-oriented Fairness [WWW21b]
Long-term Fairness [WSDM21]
Explainable Fairness [SIGIR22a, SIGIR20a]
Federated Fairness [RecSys22b]
Group-wise Fairness [RecSys17]
Fairness-Utility Relationship [WSDM22b]
Popularity Bias [CIKM21b]
Echo Chamber [SIGIR20b]
Bias and Fairness of LLMs [AACL22]

Federated Privacy [SIGIR21b, RecSys22b]
Adversarial Privacy [SIGIR21a]

Causal Robustness [CIKM22a, ICTIR23, TORS23, JCDL22]
Evaluation of Robustness [WSDM22c]

RUTGERS

Our Research Landscape – Application

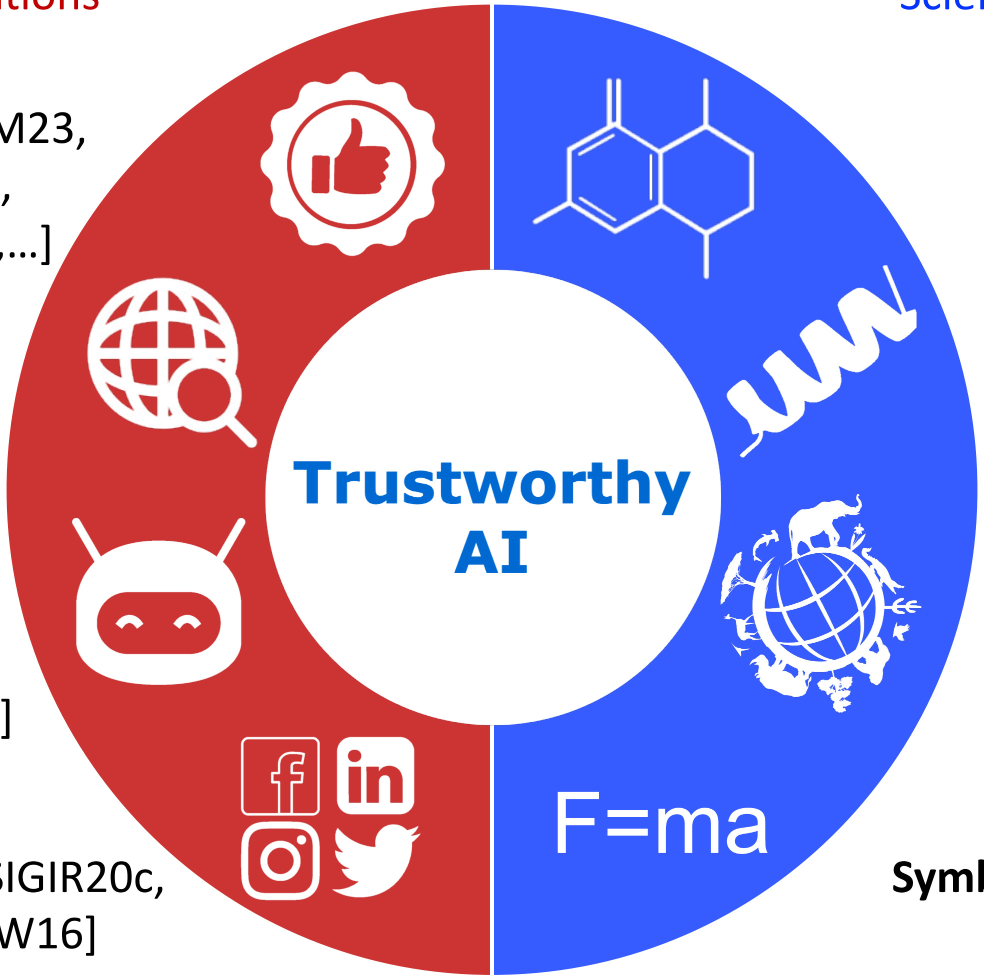
Human-oriented Applications

Recommender system [SIGIR23, WSDM23, CIKM23, WWW22b, RecSys22a, ACL22, WSDM22a-c, WWW21a-b, SIGIR21a-d,…]

Search engines [CIKM19a, TOIS19, CIKM18, SIGIR17]

QA and Dialog System [EMNLP22a, EMNLP22b, CIKM21b, SIGIR21c, SIGIR19c, CIKM19a, CIKM19b, CIKM18]

Economic and E-commerce Systems [SIGIR20c, WWW19a, WWW19b, WSDM17, WWW16]



Science-oriented Applications

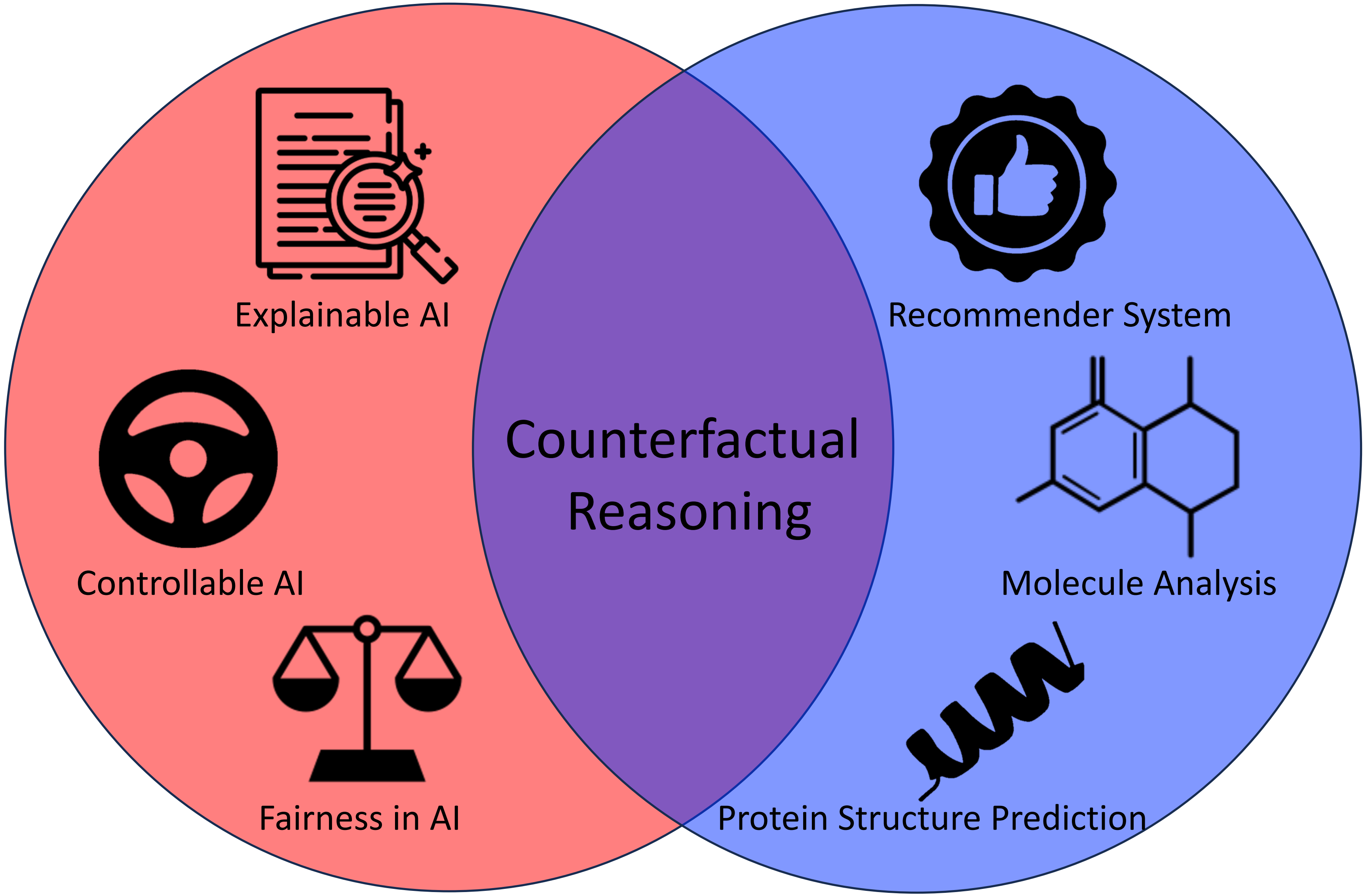
Molecule Analysis [WWW22a]

Protein Structure Prediction [KDD23]

Biodiversity Preservation [COLING20]

Symbolic Physical Rule Discovery [ICML22]

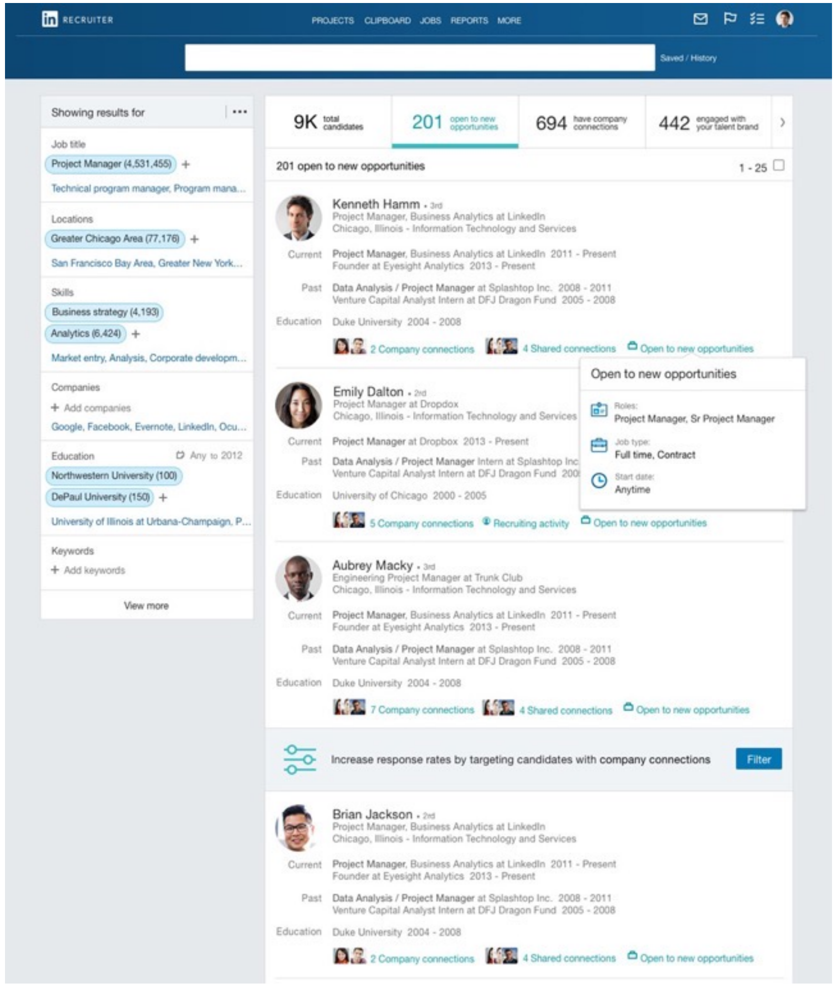
One Theme that Connects Many Dots



Multiple Trustworthy Perspectives **Multiple Application Domains**

RUTGERS

Why Trustworthy AI



Example of Human-oriented Application

Example of Science-oriented Application

RUTGERS

Example: Resume Ranking and Recommendation

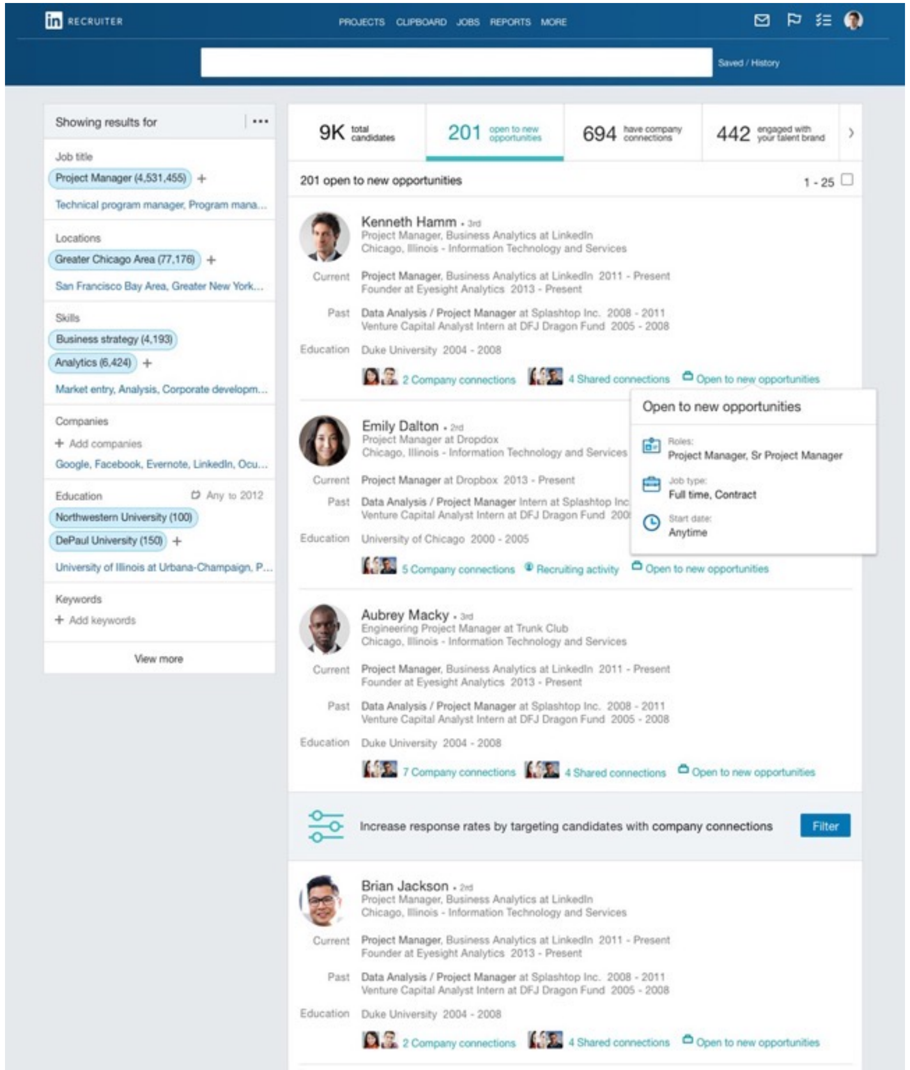


Figure 1: A (mocked) screenshot from the LinkedIn Recruiter (credit to [1])

Background: HR may use automated tools such as LinkedIn for ranking candidates due to too many applicants

Problem:
From recruiter's perspective:
Why this candidate is a better fit than another?

From applicant's perspective:
Why should I trust the algorithm?
Why should my career be decided by a machine?

To answer these **WHY** questions, we need Explainable AI

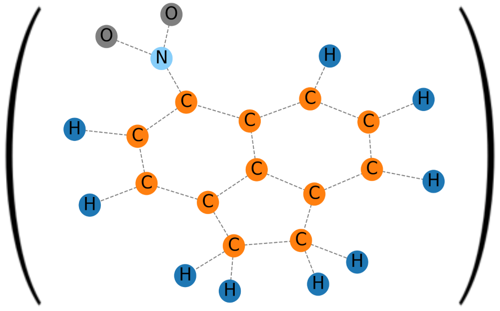
https://engineering.linkedin.com/blog/2019/04/ai-behind-linkedin-recruiter-search-and-recommendation-systems

8

RUTGERS

Example: Explainable AI for Science

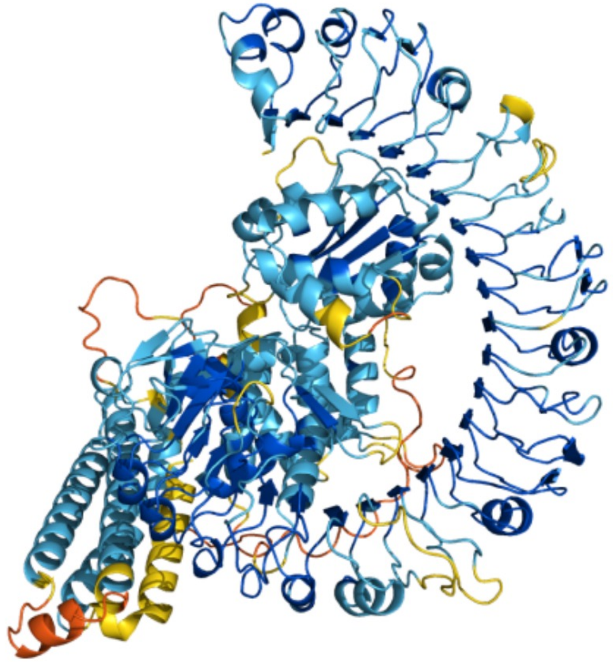
- AI for Drug Discovery
  - Molecule Property Prediction

NN  = Soluble? Toxic? Crosses the Blood Brain Barrier (BBB)?

- Protein Structure Prediction

```
MAGELVSFAVNKLWDLLSHEYTLFQGVEDQVAELKSDLNLLKSFLKDADAKKH
TSALVRYCVEEIKDIVYDAEDVLETFVQKEKLGTTSGIRKHIKRLTCIVPDRR
EIALYIGHVSKRITRVIRDMQSFGVQQMIVDDYMHPLRNREREIRRTFPKDNE
SGFVALEENVKKLVGYFVEEDNYQVVSITGMGGLGKTTLARQVFNHDMVTKKF
DKLAWVSVSQDFTLKNVWQNILGDLKPKEEETKEEEKKILEMTEYTLQRELYQ
LLEMSKSLIVLDDIWKKEDWEVIKPIFPPTKGWKLLLTSRNESIVAPTNTKYF
NFKPECLKTDDSWKLFQRIAFPINDASEFEIDEEMEKLGEKMIEHCGGLPLAI
KVLGGMLAEKYTSHDWRRLSENIGSHLVGGRTNFNDDNNNSCNYVLSLSFEEL
PSYLKHCFLYLAHFPEDYEIKVENLSYYWAAEEIFQPRHYDGEIIRDVGDVYI
EELVRRNMVISERDVKTSRFETCHLHDMMREVCLLKAKEENFLQITSNPPSTA
NFQSTVTSRRLVYQYPTTLHVEKDINNPKL…
```

AlphaFold



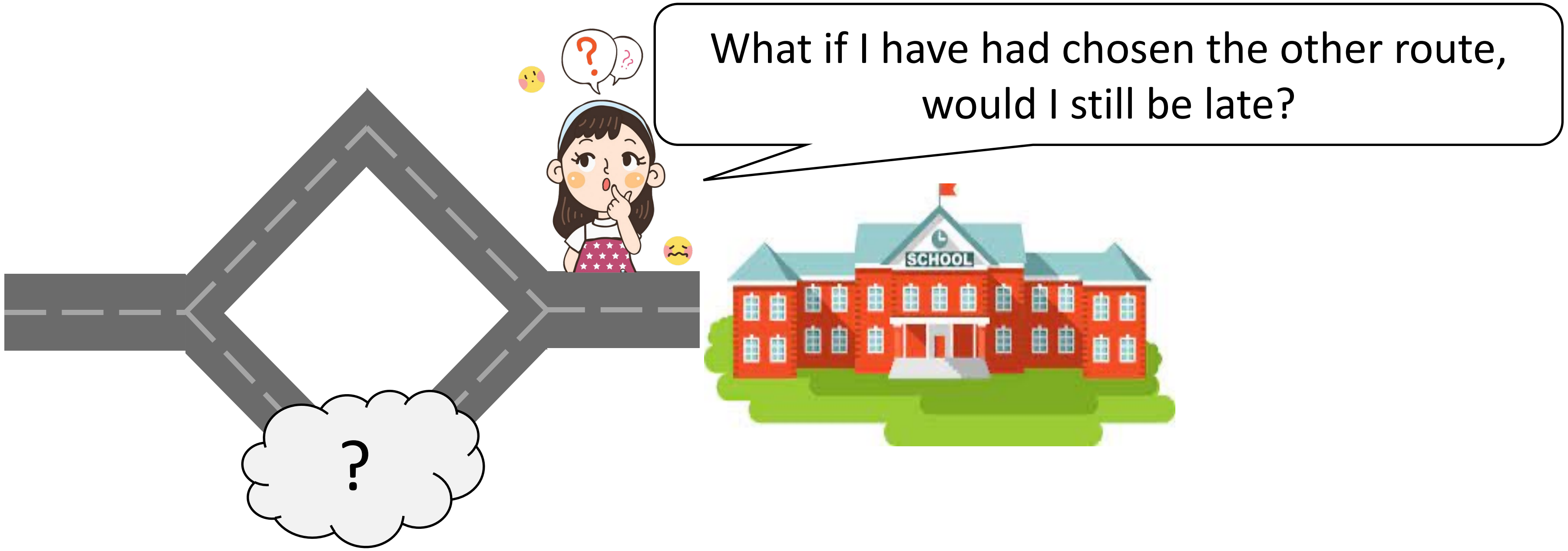
Trustworthy AI for Human

Counterfactual Reasoning and Counterfactual Explanation

[1] J Tan, S Xu, Y Ge, Y Li, X Chen, and **Y Zhang**. "Counterfactual explainable recommendation." In CIKM 2021.
[2] J Tan, Y Ge, Y Zhu, Y Xia, J Luo, J Ji, and **Y Zhang**. "User-Controllable Recommendation via Counterfactual Retrospective and Prospective Explanations." In ECAI 2023.
[3] Y Ge, J Tan, Y Zhu, Y Xia, J Luo, S Liu, Z Fu, S Geng, Z Li, and **Y Zhang**. "Explainable fairness in recommendation." In SIGIR 2022.

Counterfactual Reasoning

- Counterfactual Reasoning: the “What if” Question
  - What if something that did not happen happened?
  - What if something that happened did not happen?

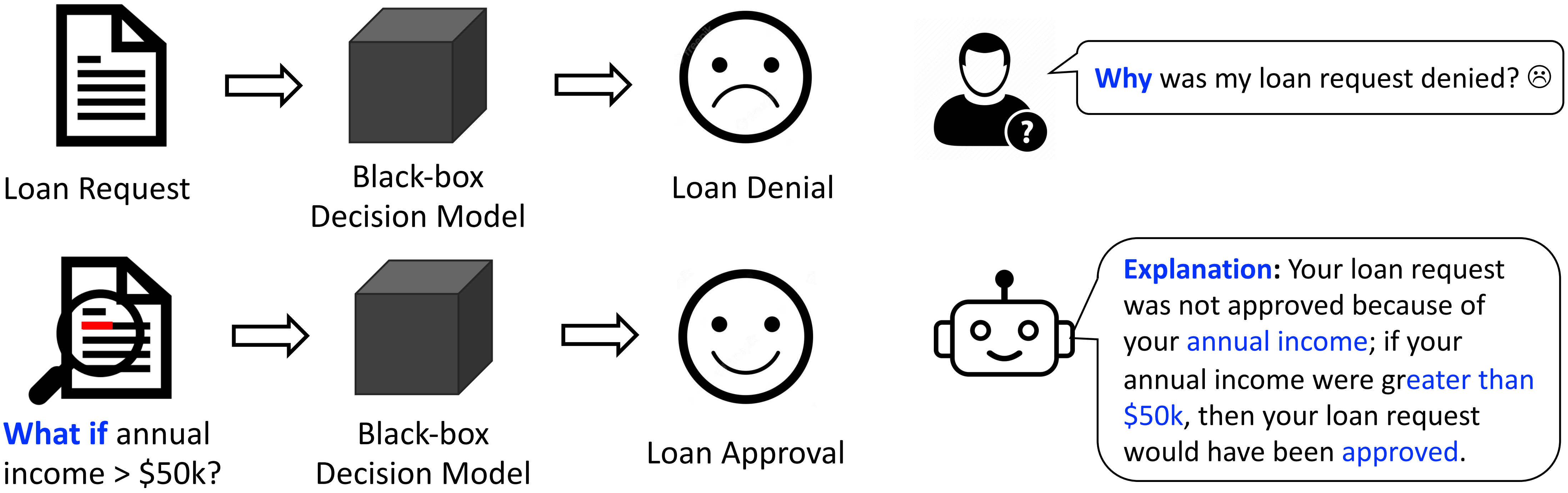


- Counterfactual reasoning shows human’s pursuit of causal relationships

RUTGERS

Counterfactual Explanation

- Explanations based on Counterfactual Reasoning



Associative Explanation vs. Causal Explanation

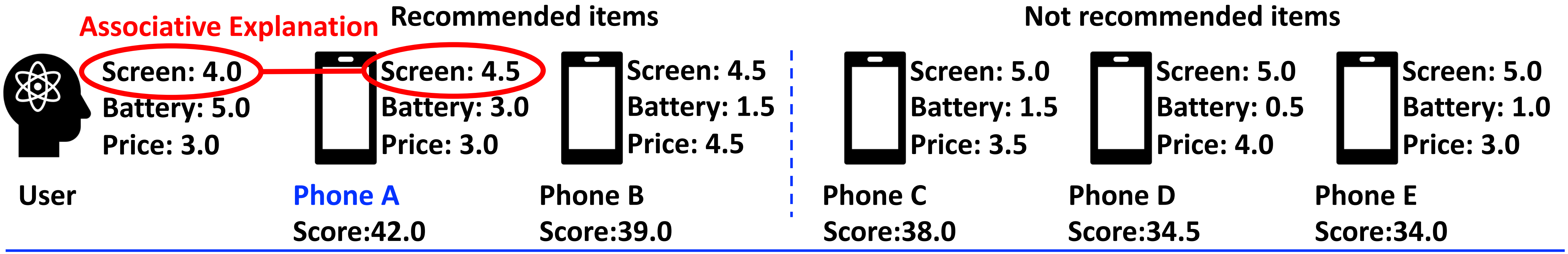
- Counterfactual Explanation is a type of Causal Explanation

	Recommended items		Not recommended items		
Screen: 4.0 Battery: 5.0 Price: 3.0	Screen: 4.5 Battery: 3.0 Price: 3.0	Screen: 4.5 Battery: 1.5 Price: 4.5	Screen: 5.0 Battery: 1.5 Price: 3.5	Screen: 5.0 Battery: 0.5 Price: 4.0	Screen: 5.0 Battery: 1.0 Price: 3.0
User	**Phone A** **Score:42.0**	**Phone B** **Score:39.0**	**Phone C** **Score:38.0**	**Phone D** **Score:34.5**	**Phone E** **Score:34.0**

[1] J Tan, S Xu, Y Ge, Y Li, X Chen, and **Y Zhang**. "Counterfactual explainable recommendation." In CIKM 2021.

Associative Explanation vs. Causal Explanation

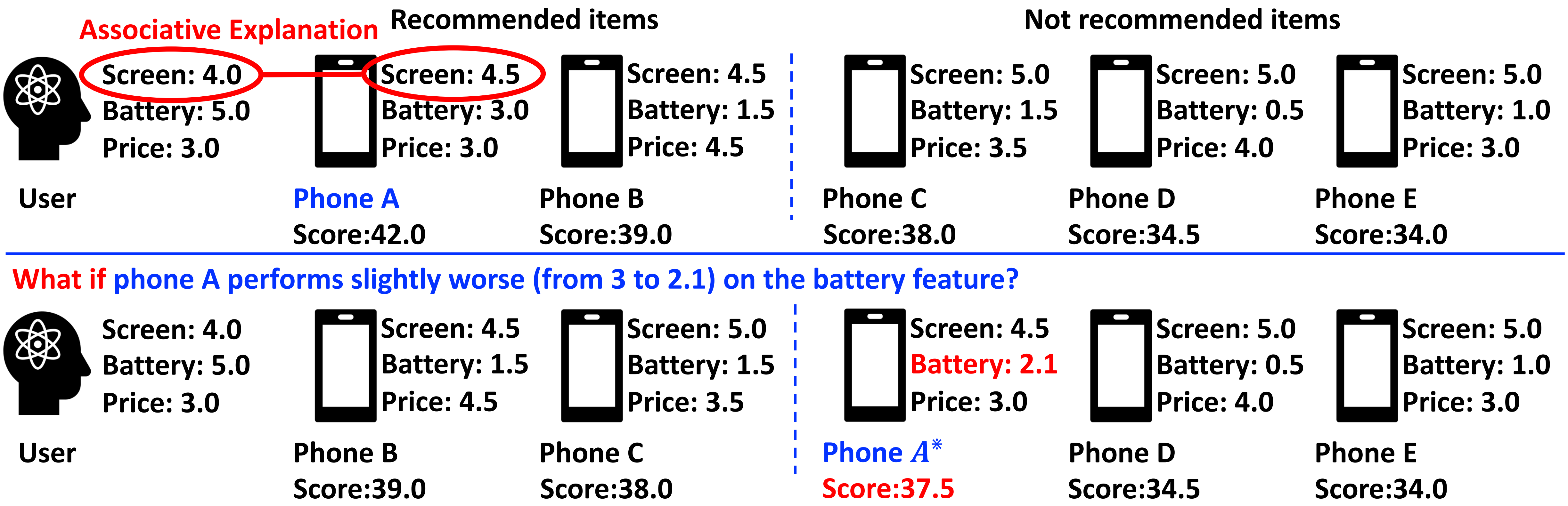
- Counterfactual Explanation is a type of Causal Explanation



[1] J Tan, S Xu, Y Ge, Y Li, X Chen, and **Y Zhang**. "Counterfactual explainable recommendation." In CIKM 2021.

Associative Explanation vs. Causal Explanation

- Counterfactual Explanation is a type of Causal Explanation



[1] J Tan, S Xu, Y Ge, Y Li, X Chen, and **Y Zhang**. "Counterfactual explainable recommendation." In CIKM 2021.

Associative Explanation vs. Causal Explanation

- Counterfactual Explanation is a type of Causal Explanation

Associative Explanation

	Recommended items		Not recommended items		
Screen: 4.0 Battery: 5.0 Price: 3.0	Screen: 4.5 Battery: 3.0 Price: 3.0	Screen: 4.5 Battery: 1.5 Price: 4.5	Screen: 5.0 Battery: 1.5 Price: 3.5	Screen: 5.0 Battery: 0.5 Price: 4.0	Screen: 5.0 Battery: 1.0 Price: 3.0
User	Phone A Score:42.0	Phone B Score:39.0	Phone C Score:38.0	Phone D Score:34.5	Phone E Score:34.0

What if phone A performs slightly worse (from 3 to 2.1) on the battery feature?

Screen: 4.0 Battery: 5.0 Price: 3.0	Screen: 4.5 Battery: 1.5 Price: 4.5	Screen: 5.0 Battery: 1.5 Price: 3.5	Screen: 4.5 Battery: 2.1 Price: 3.0	Screen: 5.0 Battery: 0.5 Price: 4.0	Screen: 5.0 Battery: 1.0 Price: 3.0
User	Phone B Score:39.0	Phone C Score:38.0	Phone A^* Score:37.5	Phone D Score:34.5	Phone E Score:34.0

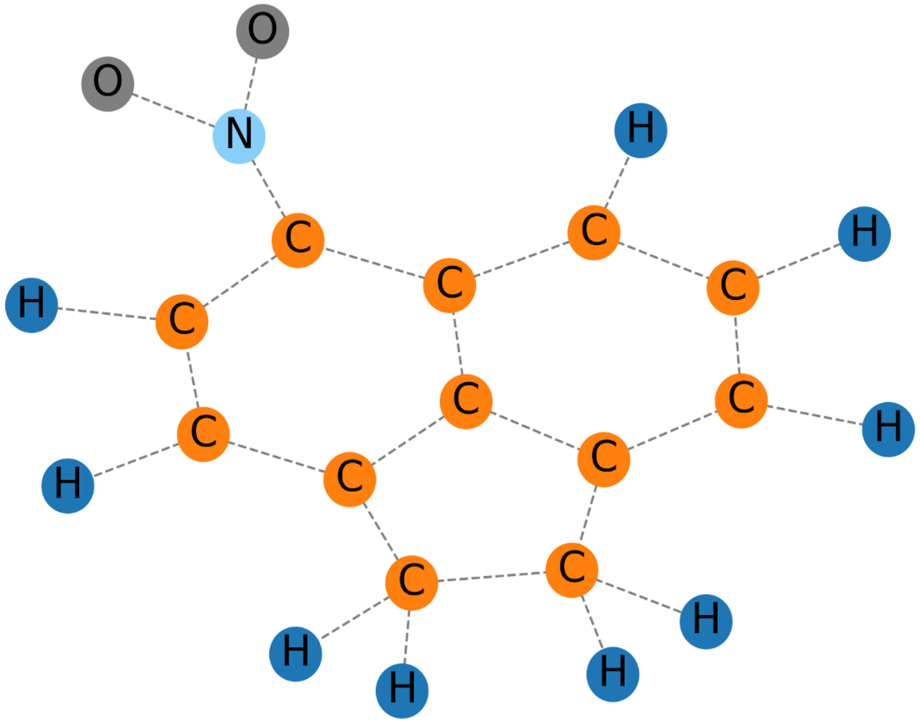
Counterfactual Explanation

If the item had been slightly worse on [feature], then it would not have been recommended at all.

[1] J Tan, S Xu, Y Ge, Y Li, X Chen, and **Y Zhang**. "Counterfactual explainable recommendation." In CIKM 2021.

RUTGERS

Counterfactual Explanation on Graphs



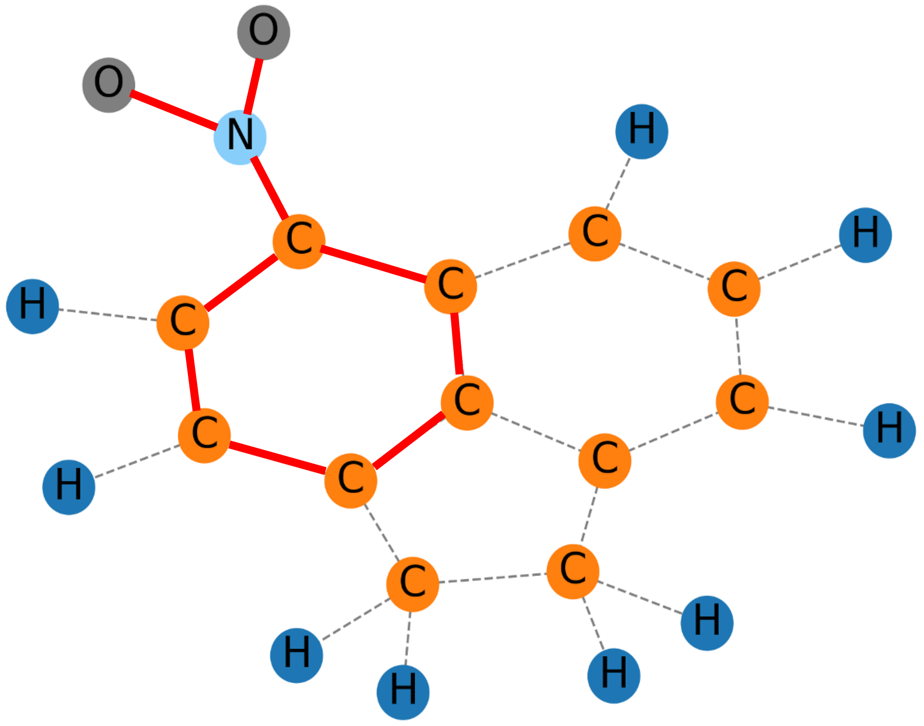
A mutagenic molecule

Why is this molecule toxic (mutagenic)?

[2] J Tan, S Geng, Z Fu, Y Ge, S Xu, Y Li, and **Y Zhang**. "Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning." In WWW 2022.

RUTGERS

Counterfactual Explanation on Graphs



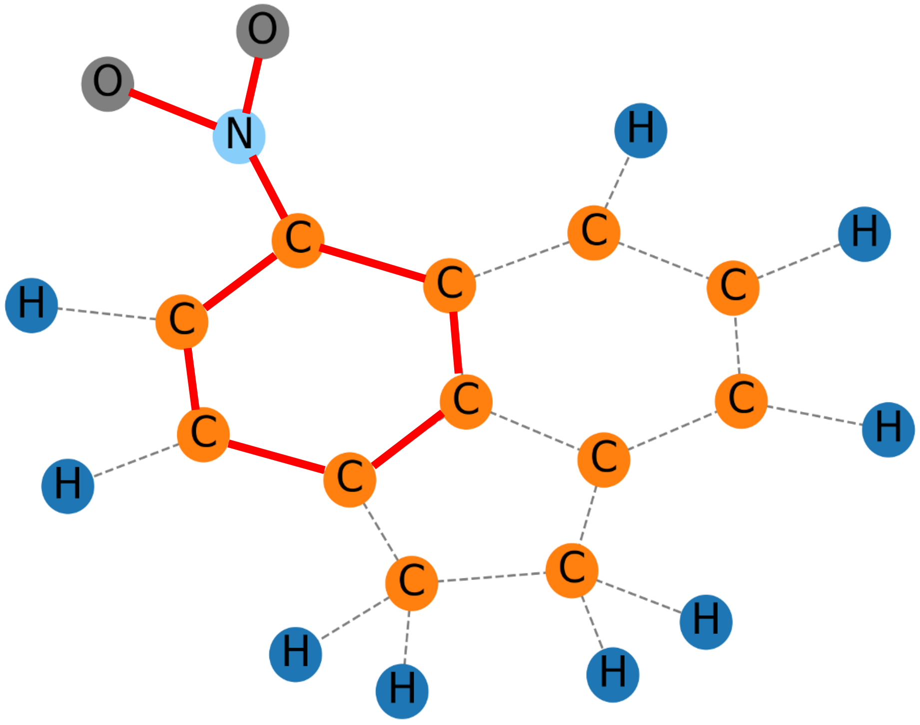
A mutagenic molecule

Why is this molecule toxic (mutagenic)?

Explanation: The **Nitrobenzene** structure

[2] J Tan, S Geng, Z Fu, Y Ge, S Xu, Y Li, and **Y Zhang**. "Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning." In WWW 2022.

Counterfactual Explanation on Graphs

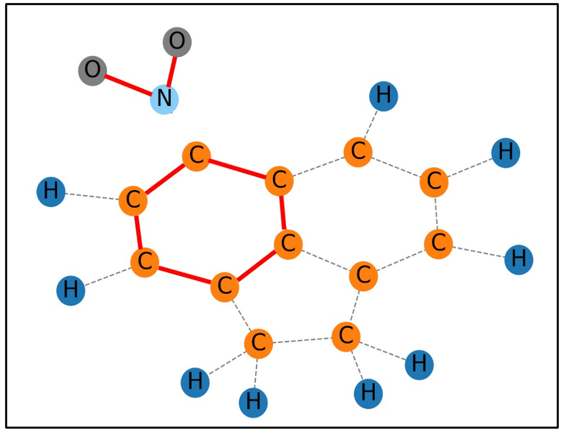


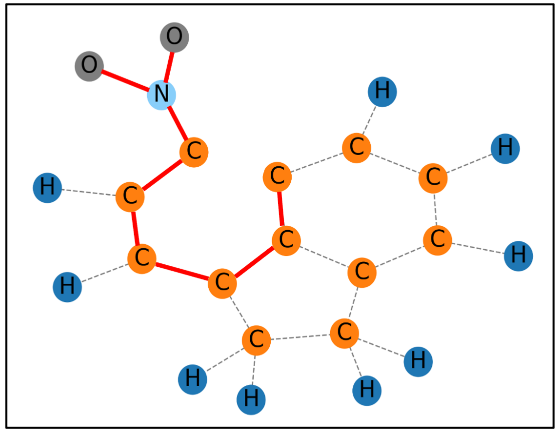
A mutagenic molecule

Why is this molecule toxic (mutagenic)?

Explanation: The **Nitrobenzene** structure.

If the Nitrobenzene structure were broken, then the molecule would not have been toxic at all.





[2] J Tan, S Geng, Z Fu, Y Ge, S Xu, Y Li, and **Y Zhang**. "Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning." In WWW 2022.

Simple and Effective Explanations (CIKM’21)

What is a good explanation?

How to find the explanation?

How to evaluate the explanation?

[1] J Tan, S Xu, Y Ge, Y Li, X Chen, and **Y Zhang**. "Counterfactual explainable recommendation." In CIKM 2021.

What is a Good Explanation?

- A good explanation is Simple and Effective

Occam's Razor Principle for Explainable AI [1]:
When trying to explain a phenomenon, if two explanations are equally effective, then we prefer the simpler one.

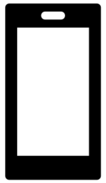
[1] J Tan, S Xu, Y Ge, Y Li, X Chen, and **Y Zhang**. "Counterfactual explainable recommendation." In CIKM 2021.

How to define Simplicity and Effectiveness?

- Counterfactual Explanation as Intervention Vector
  - Item Representation Vector

$Z =$

Screen	Battery	Price
4.5	3.0	3.0



Screen: 4.5
Battery: 3.0
Price: 3.0

  - Explanation as an Intervention Vector

$\Delta =$

Screen	Battery	Price
0	-0.9	0

  - Item Representation after Counterfactual Intervention

$$Z' = Z + \Delta$$

[1] J Tan, S Xu, Y Ge, Y Li, X Chen, and **Y Zhang**. "Counterfactual explainable recommendation." In CIKM 2021.

RUTGERS

How to define Simplicity and Effectiveness?

To define Simplicity:

Explanation Complexity

$$C(\Delta) = \underbrace{\gamma||\Delta||_0}_{} + \underbrace{||\Delta||_2^2}_{}$$

# of non-zeros in Δ, i.e., number of features we need to change

Square of Δ, i.e., the degree of change we need to apply on the features

To define Effectiveness:

Explanation Strength

$$S(\Delta) = \underbrace{s_{i,j} - s_{i,j_\Delta}}_{}$$

Change of the item's ranking score before and after applying the interventions

Simplicity means low complexity: change **as few features as possible** and the change should be **as small as possible**

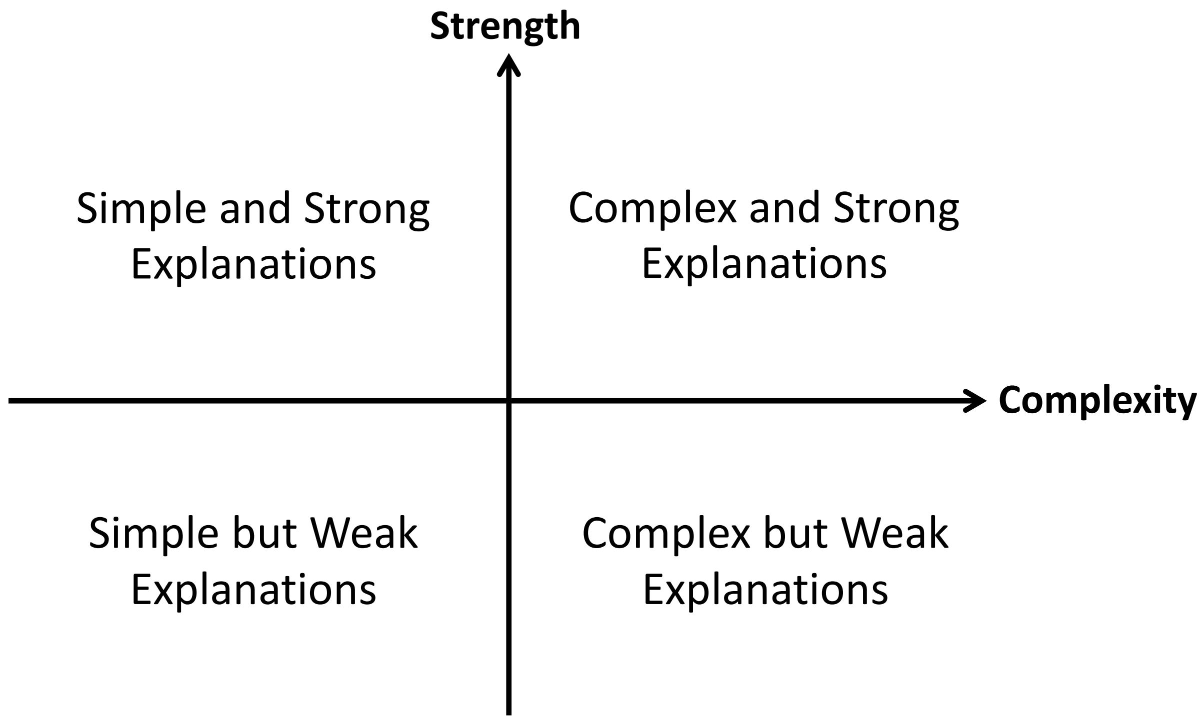
Effectiveness means high strength: item's ranking score should be reduced **large enough to be removed from the top-K list**

User	Phone B	Phone C	Phone A^*	Phone D	Phone E
Screen: 4.0	**Screen: 4.5**	**Screen: 5.0**	**Screen: 4.5**	**Screen: 5.0**	**Screen: 5.0**
Battery: 5.0	**Battery: 1.5**	**Battery: 1.5**	**Battery: 2.1**	**Battery: 0.5**	**Battery: 1.0**
Price: 3.0	**Price: 4.5**	**Price: 3.5**	**Price: 3.0**	**Price: 4.0**	**Price: 3.0**
	Score:39.0	**Score:38.0**	**Score:37.5**	**Score:34.5**	**Score:34.0**

Counterfactual Explanation: If the item had been slightly worse on [feature], then it would not have been recommended at all.

Complexity vs. Strength

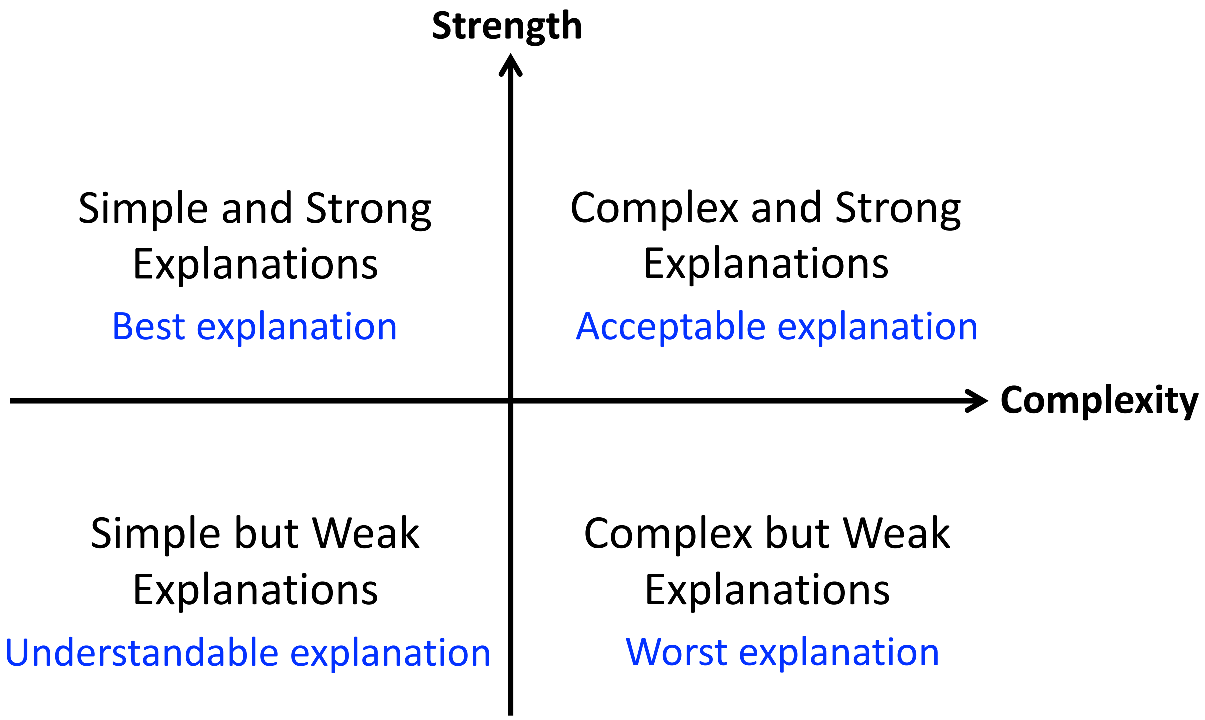
- Two Orthogonal Dimensions
  - Complex explanations may not be strong, Simple explanations may not be weak
  - There exist complex but weak explanations, or simple and strong explanations



RUTGERS

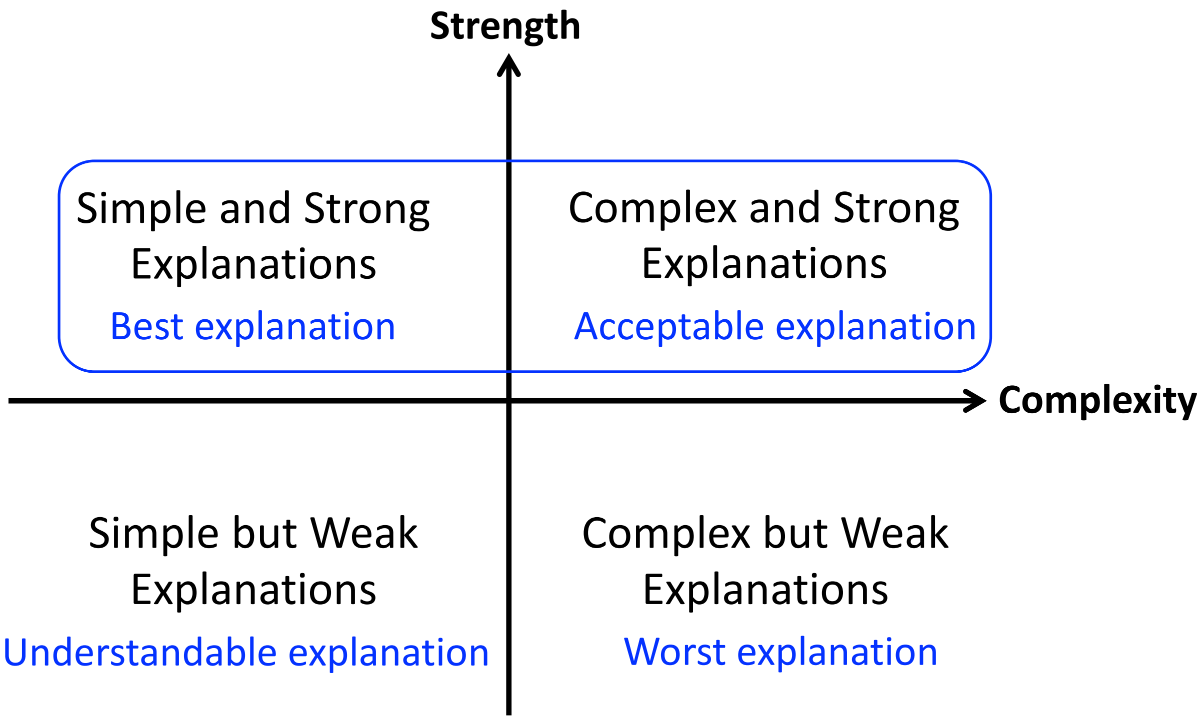
Complexity vs. Strength

- Two Orthogonal Dimensions
  - Complex explanations may not be strong, Simple explanations may not be weak
  - There exist complex but weak explanations, or simple and strong explanations



Complexity vs. Strength

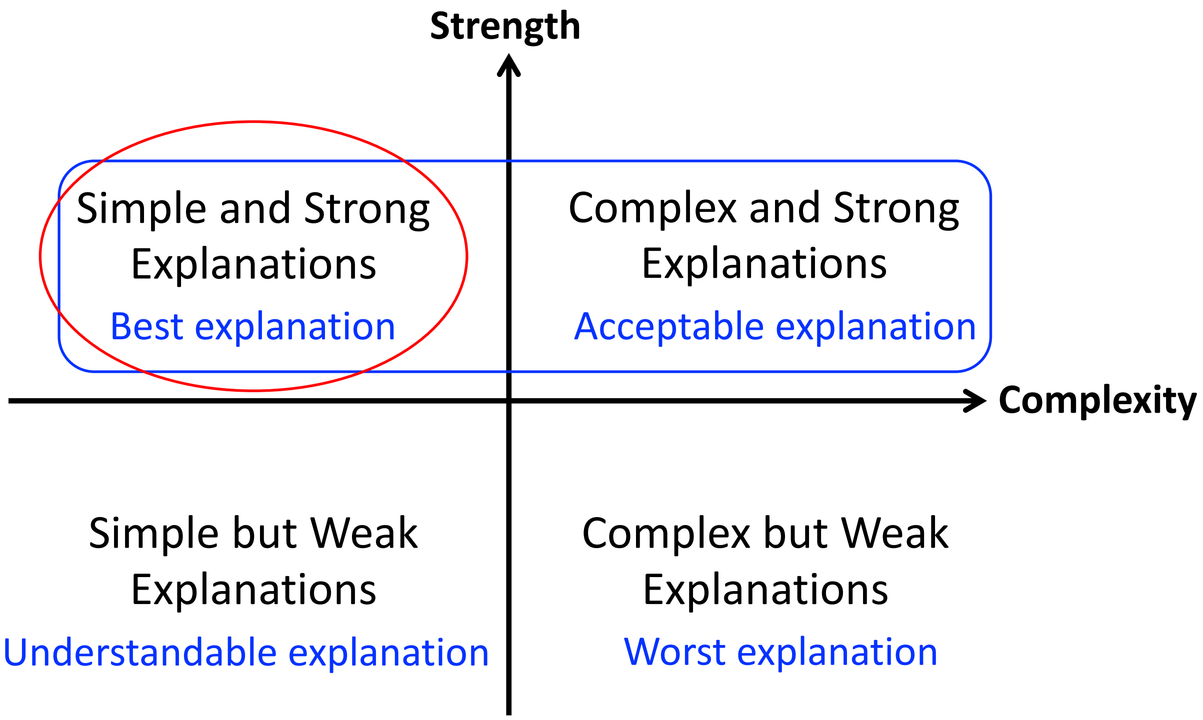
- Two Orthogonal Dimensions
  - Complex explanations may not be strong, Simple explanations may not be weak
  - There exist complex but weak explanations, or simple and strong explanations



We need to guarantee the explanation is strong in the first place (High Strength)

Complexity vs. Strength

- Two Orthogonal Dimensions
  - Complex explanations may not be strong, Simple explanations may not be weak
  - There exist complex but weak explanations, or simple and strong explanations



We need to guarantee the explanation is strong in the first place (High Strength)

Given that, seek for simple explanations (Low Complexity)

RUTGERS

How to Learn Counterfactual Explanations?

- A Counterfactual Constrained Learning Framework
  - Black-box Prediction Model $s_{ij} = f(Y_i, Z_j|\Theta)$
    - s_{ij}: algorithm's predicted score for user u_i on item v_j

$$\begin{aligned} &\text{minimize Explanation Complexity} \\ &\quad \text{s.t. Explanation is Strong Enough} \end{aligned} \quad \Rightarrow \quad \begin{aligned} &\underset{\Delta}{\text{minimize}} \ \|\Delta\|_2^2 + \gamma\|\Delta\|_0 \\ &\quad \text{s.t.} \ s_{i,j_\Delta} \le s_{i,j_{K+1}} \end{aligned}$$

Seek for simple (low complexity) explanations constrained on that the explanation is strong enough

- $s_{i,j_\Delta} = f\left(Y_i, Z_j + \Delta \middle| \Theta\right)$: score of item v_j after applying intervention vector Δ
- $s_{i,j_{K+1}} = f(Y_i, Z_{j_{K+1}}|\Theta)$: score of the item that was originally ranked at position $K+1$
- The framework can be applied on any black-box prediction model

How to Learn Counterfactual Explanations?

- A Counterfactual Constrained Learning Framework
  - Black-box Prediction Model $s_{ij} = f(Y_i, Z_j | \Theta)$
    - s_{ij}: algorithm's predicted score for user u_i on item v_j

$$\text{minimize Explanation Complexity} \quad \text{s.t. Explanation is Strong Enough}$$

$\Rightarrow$

$$\underset{\Delta}{\text{minimize}} \ \|\Delta\|_2^2 + \gamma\|\Delta\|_0 \quad \text{s.t.} \ s_{i,j_\Delta} \le s_{i,j_{K+1}}$$

Relaxed optimization with Lagrange multiplier:

$$\underset{\Delta}{\text{minimize}} \ \|\Delta\|_2^2 + \gamma\|\Delta\|_1 + \lambda L(s_{i,j_\Delta}, s_{i,j_{K+1}})$$

$$\text{where:} \quad L\left(s_{i,j_\Delta}, s_{i,j_{K+1}}\right) = \max(0, \alpha + s_{i,j_\Delta} - s_{i,j_{K+1}})$$

(0-norm $\|\Delta\|_0$ is replaced with 1-norm $\|\Delta\|_1$: optimizable and gives sparsity)

RUTGERS

How to Evaluate Counterfactual Explanations?

Sufficiency and Necessity:

$S \Rightarrow N$: S is a sufficient condition for N

$\neg N \Rightarrow \neg S$: N is a necessary condition for S

Two metrics for evaluating Counterfactual Explanations

Probability of Necessity (PN)

Probability of Sufficiency (PS)

Probability of Necessity (PN)

- Counterfactual Question:
  - If the explanation feature had not existed, would the item still be recommended?

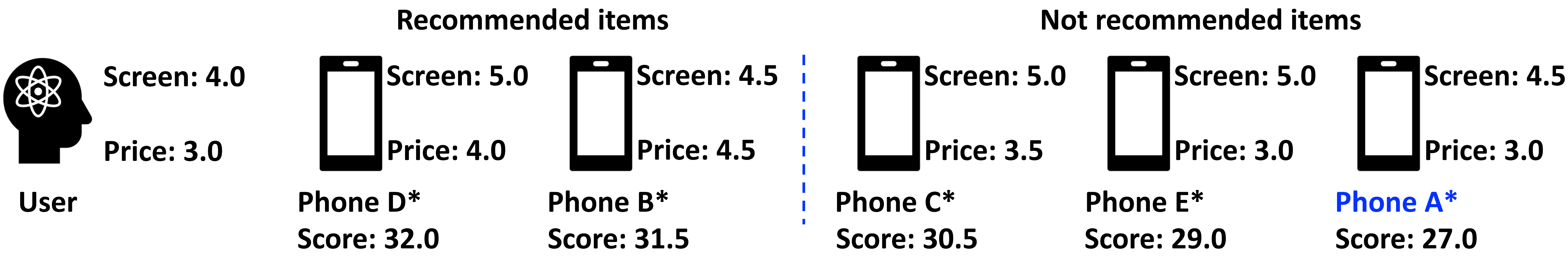
Recommended items | **Not recommended items**

User	Phone A	Phone B	Phone C	Phone D	Phone E
Screen: 4.0	Screen: 4.5	Screen: 4.5	Screen: 5.0	Screen: 5.0	Screen: 5.0
~~Battery: 5.0~~	~~Battery: 3.0~~	~~Battery: 1.5~~	~~Battery: 1.5~~	~~Battery: 0.5~~	~~Battery: 1.0~~
Price: 3.0	Price: 3.0	Price: 4.5	Price: 3.5	Price: 4.0	Price: 3.0
	Score: 42.0	Score: 39.0	Score: 38.0	Score: 34.5	Score: 34.0

RUTGERS

Probability of Necessity (PN)

- Counterfactual Question:
  - If the explanation feature had not existed, would the item still be recommended?
  - If the answer is NO, then it is a necessary explanation

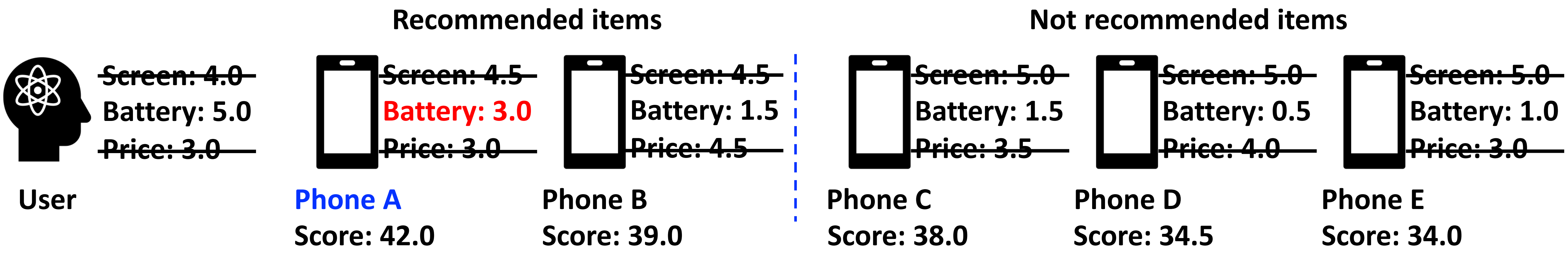


$$\mathrm{PN} = \frac{\sum_{u_i \in \mathcal{U}} \sum_{v_j \in R_{i,K}} \mathrm{PN}_{ij}}{\sum_{u_i \in \mathcal{U}} \sum_{v_j \in R_{i,K}} I(\mathcal{A}_{ij} \neq \emptyset)}, \text{ where } \mathrm{PN}_{ij} = \begin{cases} 1, & \text{if } v_j^* \notin R_{i,K}^* \\ 0, & \text{else} \end{cases}$$

PN: Percentage of explanation that satisfy the above necessity criterion

Probability of Sufficiency (PS)

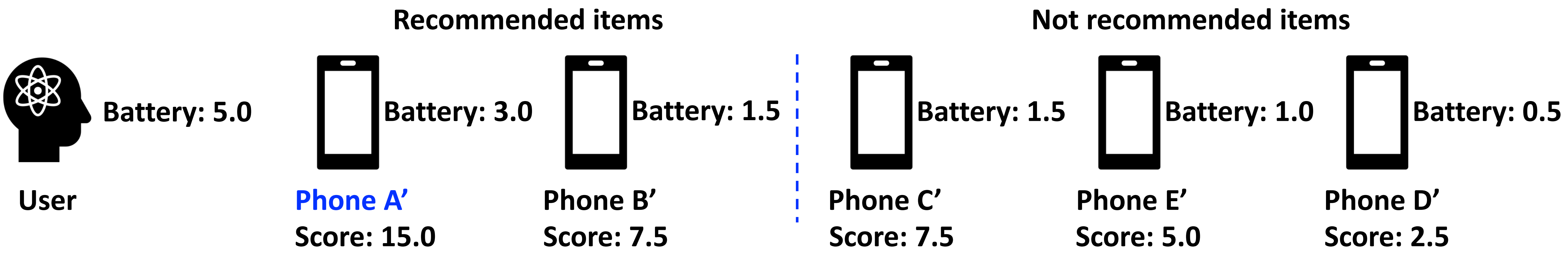
- Counterfactual Question:
  - If the explanation feature were the only feature, would the item still be recommended?



RUTGERS

Probability of Sufficiency (PS)

- Counterfactual Question:
  - If the explanation feature were the only feature, would the item still be recommended?
  - If the answer is YES, then it is a sufficient explanation



$$\mathrm{PS} = \frac{\sum_{u_i \in \mathcal{U}} \sum_{v_j \in R_{i,K}} \mathrm{PS}_{ij}}{\sum_{u_i \in \mathcal{U}} \sum_{v_j \in R_{i,K}} I(\mathcal{A}_{ij} \neq \emptyset)}, \text{ where } \mathrm{PS}_{ij} = \begin{cases} 1, & \text{if } v'_j \in R'_{i,K} \\ 0, & \text{else} \end{cases}$$

PS: Percentage of explanation that satisfy the above sufficiency criterion

Evaluation of Counterfactual Explanation

- Counterfactual Explanations better than Associative Explanations

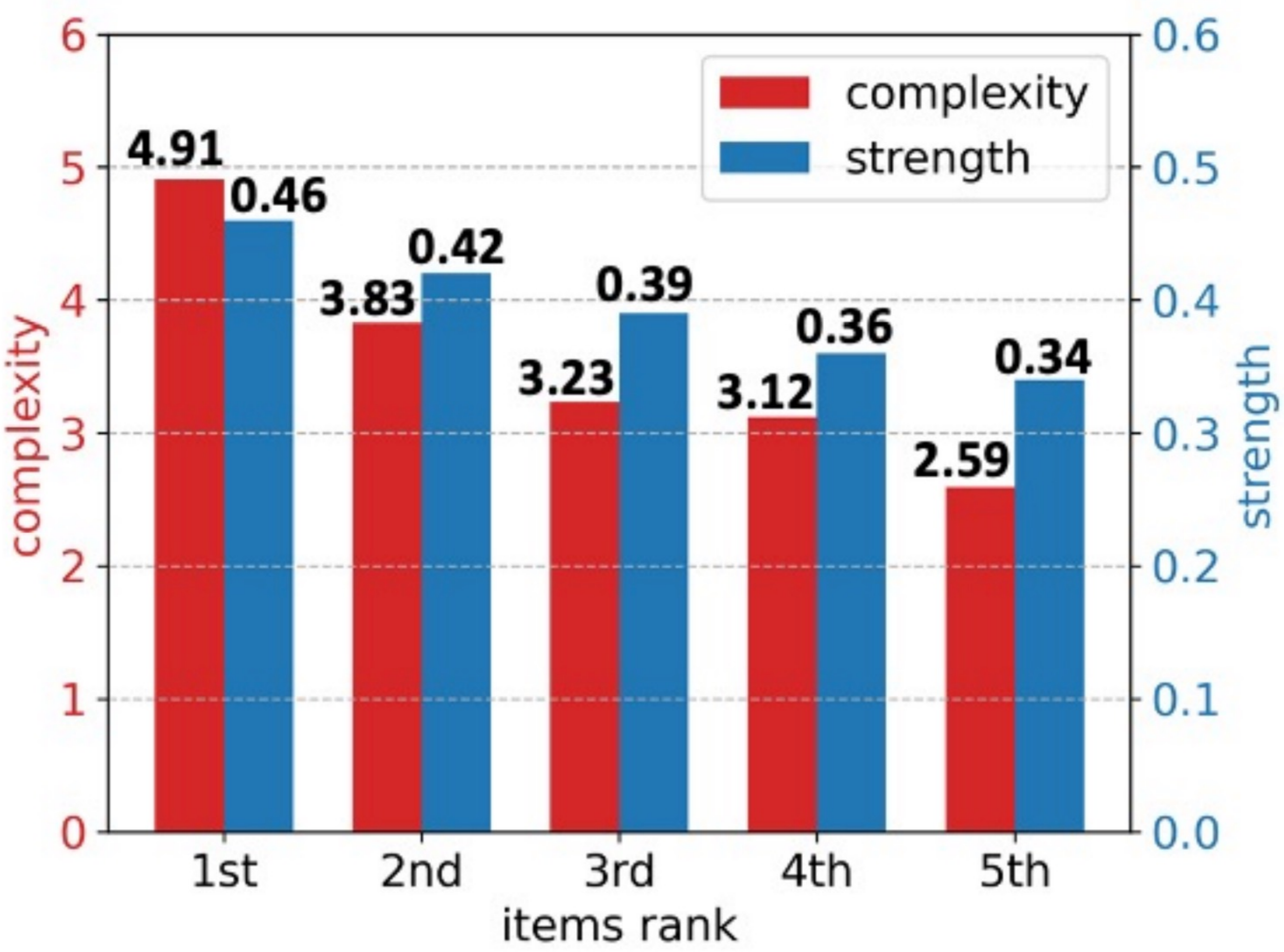
	Single Aspect Explanation														
	Electronic			Cell Phones			Kindle Store			CDs and Vinyl			Yelp		
	PN%	PS%	F_{NS}%	PN%	PS%	F_{NS}%	PN%	PS%	F_{NS}%	PN%	PS%	F_{NS}%	PN%	PS%	F_{NS}%
Random	2.05	2.10	2.07	3.39	3.50	3.44	3.16	2.75	2.94	1.58	2.03	1.78	7.52	10.68	8.82
EFM[50]	8.41	41.13	13.96	32.31	**82.09**	46.37	6.01	**73.84**	11.12	10.15	42.63	16.39	5.87	61.06	10.71
A2CF[9]	41.45	**77.60**	54.03	36.82	78.68	50.17	25.66	65.53	36.88	25.41	**84.51**	39.07	17.59	**96.92**	29.78
CountER	**65.54**	68.28	**66.83**	**74.03**	63.30	**68.25**	34.37	41.50	37.60	49.62	54.72	52.04	**65.26**	53.25	**58.64**

	Multiple Aspect Explanation														
	Electronic			Cell Phones			Kindle Store			CDs and Vinyl			Yelp		
	PN%	PS%	F_{NS}%	PN%	PS%	F_{NS}%	PN%	PS%	F_{NS}%	PN%	PS%	F_{NS}%	PN%	PS%	F_{NS}%
Random	2.24	4.90	3.08	6.25	10.13	7.73	5.80	7.80	6.65	3.22	7.65	4.53	13.84	12.92	13.36
EFM[50]	29.65	84.67	43.92	52.66	87.98	65.88	51.72	**96.42**	67.33	47.65	87.35	61.66	16.76	81.68	27.81
A2CF[9]	59.47	81.66	68.82	56.45	80.97	66.52	52.48	87.59	65.64	49.12	**91.52**	63.93	41.38	**98.28**	58.24
CountER	**97.08**	**96.24**	**96.66**	**99.52**	**98.48**	**99.00**	**64.00**	79.20	**70.79**	**80.89**	88.60	**84.57**	**99.91**	94.12	**96.93**

RUTGERS

Interesting Observation

- Top-ranked items need to be backed by stronger and more complex explanations



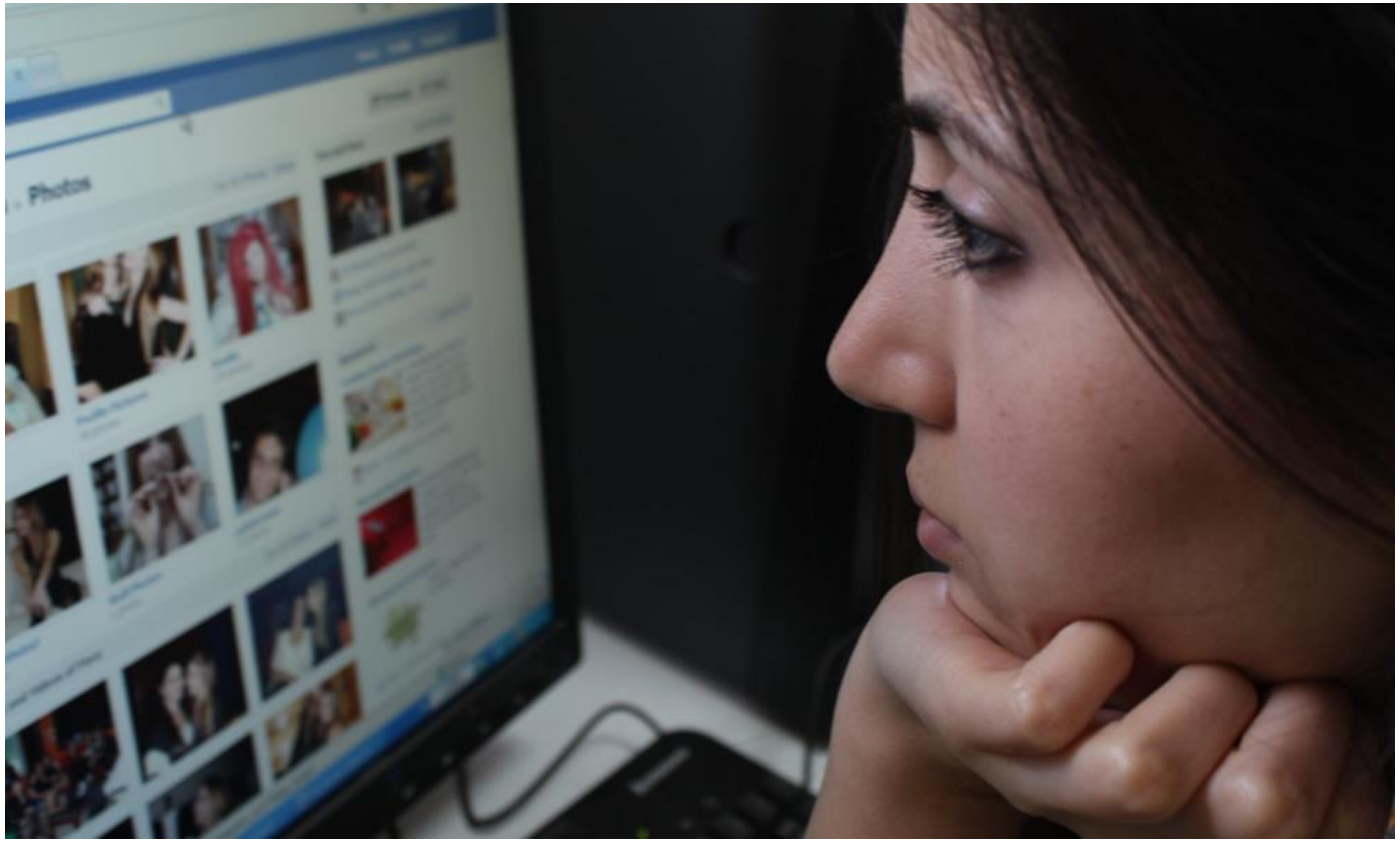
User Controllable AI (ECAI’23)

[2] J Tan, Y Ge, Y Zhu, Y Xia, J Luo, J Ji, and **Y Zhang**. "User-Controllable Recommendation via Counterfactual Retrospective and Prospective Explanations." In ECAI 2023.

RUTGERS

Towards User Controllable Recommender Systems

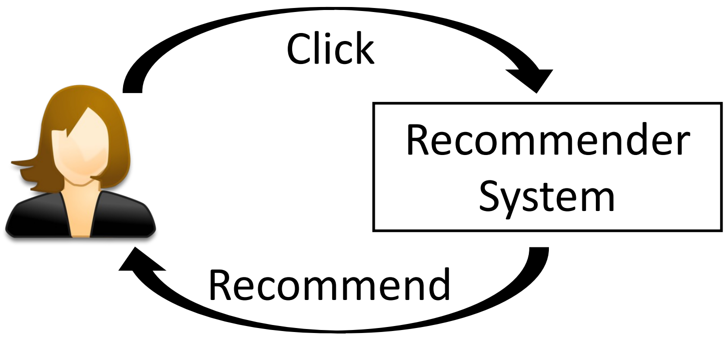
- Users almost have no control of their recommender system
  - They can only passively receive recommendations



RUTGERS

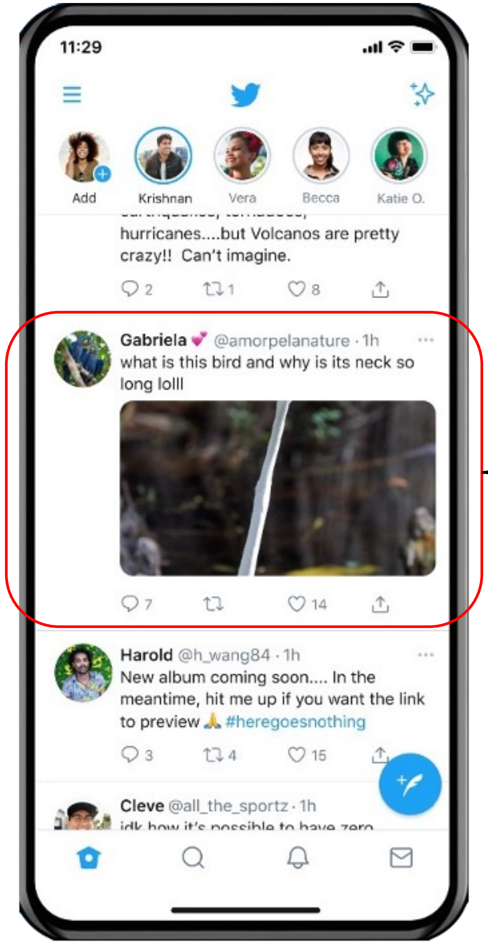
Towards User Controllable Recommender Systems

- Users almost have no control of their recommender system
  - They can only passively receive recommendations
- This causes many problems, e.g., echo chamber



The more you like something, the more RS will recommend similar things, and thus you like them even more.

User Control based on Counterfactual Explanations



Counterfactual Retrospective Explanation [3]

We recommend this video X because you previously ❤️videos A and B, if you did not ❤️ them, then we would not have recommended this video X.

Counterfactual Prospective Explanation [3]

If you ❤️ this video X, then we will recommend videos D and E in the future that otherwise would not be recommended.

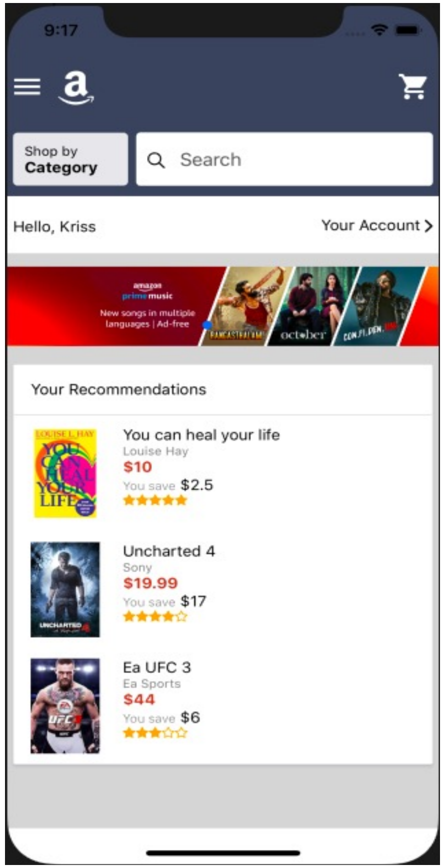
Help users know the consequences of their behaviors so that they can take informed actions.
Users can control their recommendation by invoking or revoking certain actions.

[2] J Tan, Y Ge, Y Zhu, Y Xia, J Luo, J Ji, and **Y Zhang**. "User-Controllable Recommendation via Counterfactual Retrospective and Prospective Explanations." In ECAI 2023.

Counterfactual Explainable Fairness (SIGIR'22)

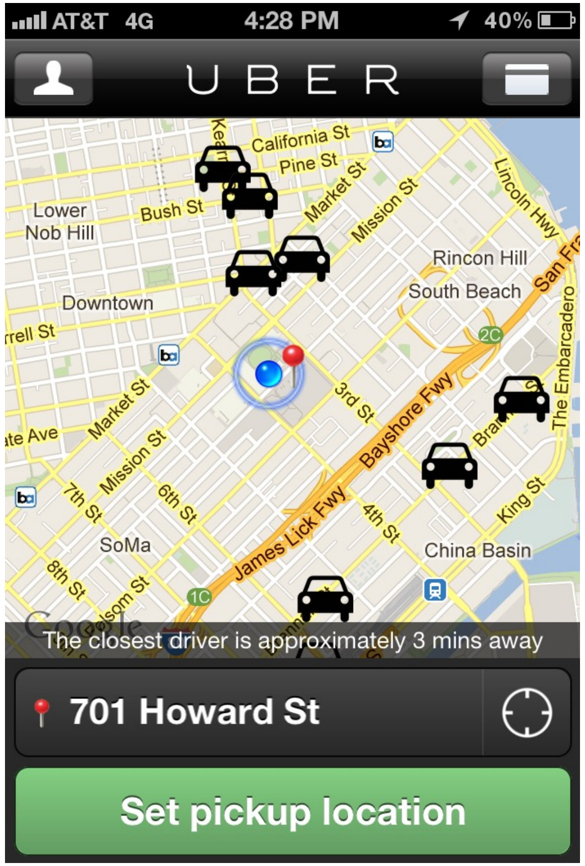
[3] Y Ge, J Tan, Y Zhu, Y Xia, J Luo, S Liu, Z Fu, S Geng, Z Li, and **Y Zhang**. "Explainable fairness in recommendation." In SIGIR 2022.

Why Fairness in RecSys? Resource is Limited

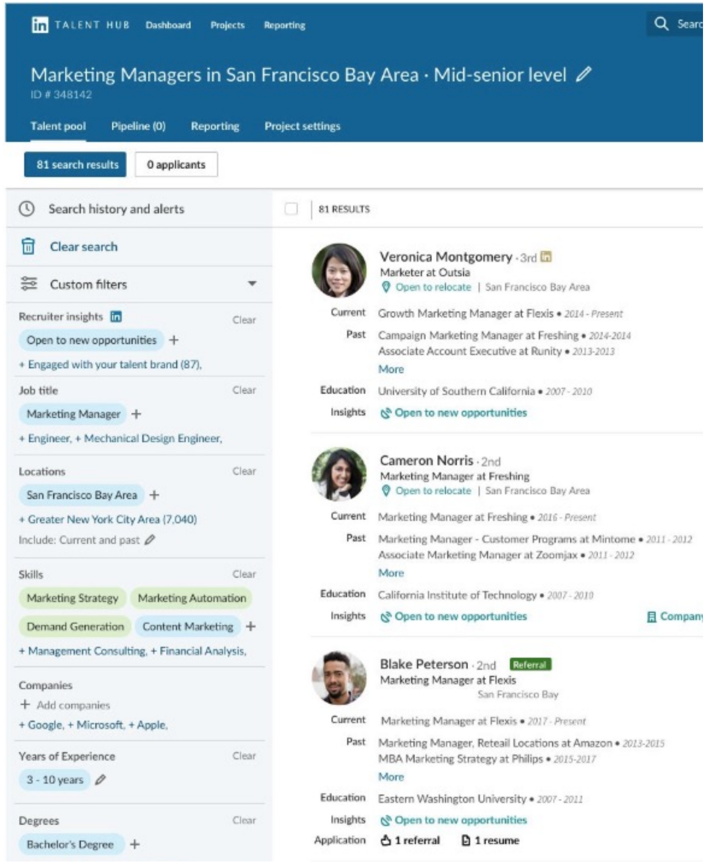


Recommendation slot positions are limited

User attention is a limited resource



Passengers are limited

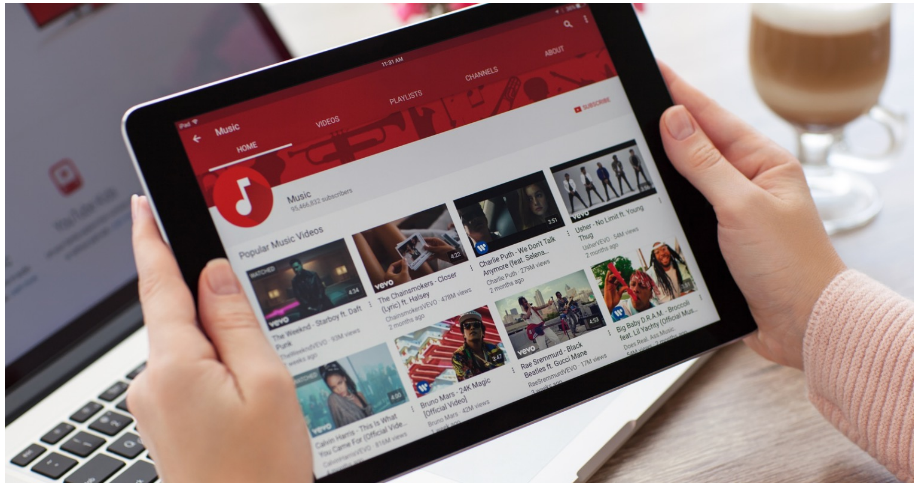


Interview opportunities are limited

Fairness and Sustainable Development

- RecSys platforms consider fairness for sustainable development

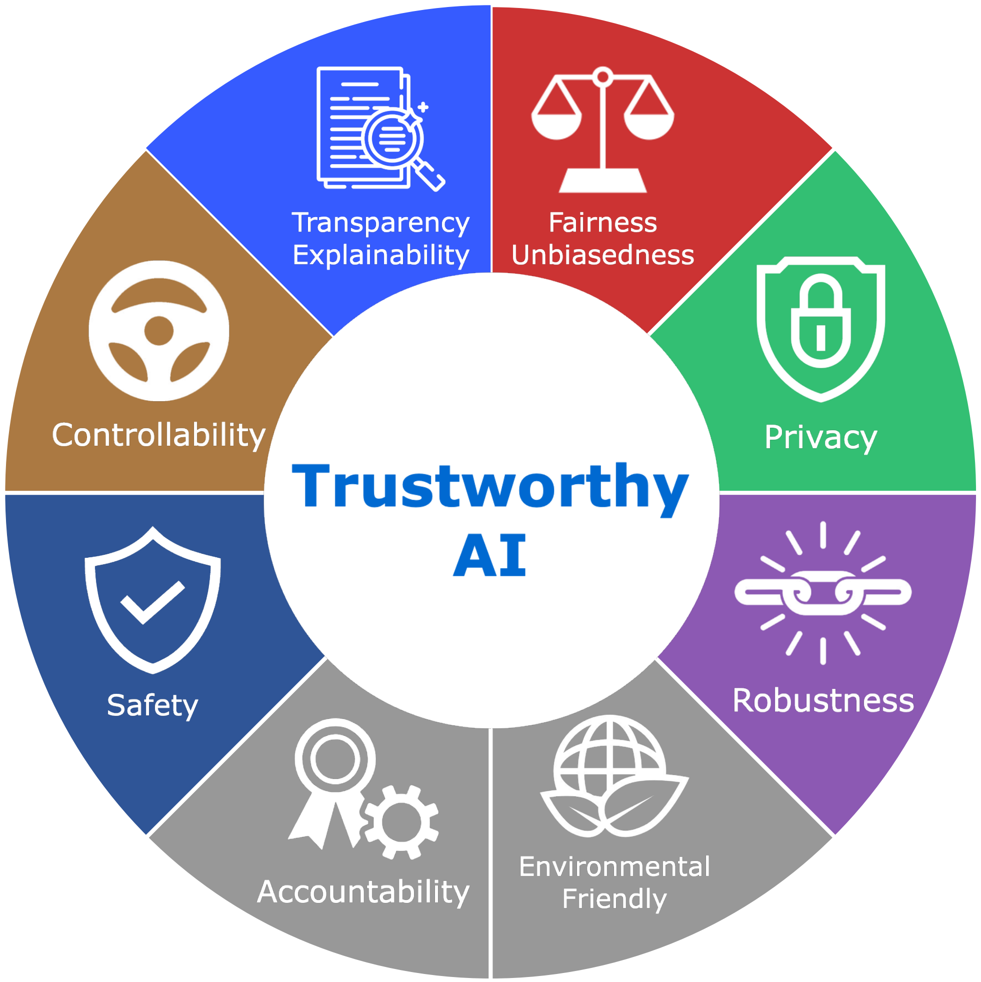
An e-commerce example
Big retailors vs. Small retailors



A social network example
Star accounts vs. Grassroot accounts

Various Types of Fairness Definitions

- Explainable Fairness based on Counterfactual Reasoning



Counterfactual Fairness [SIGIR21a]
User-oriented Fairness [WWW21b]
Long-term Fairness [WSDM21]
Explainable Fairness [SIGIR22a, SIGIR20a]
Federated Fairness [RecSys22b]
Group-wise Fairness [RecSys17]
Fairness-Utility Relationship [WSDM22b]
Popularity Bias [CIKM21b]
Echo Chamber [SIGIR20b]
Bias and Fairness of LLMs [AACL22]

Why Explainable Fairness?

- Explainable Fairness is important in Recommendation [4]
  - Hundreds, thousands or even more features
  - $y = f(F_u, F_v) = f(Lo, In, Ta, \dots, Ft, Ra, Pa, Wa, \dots)$

User ID	Item ID	Location	Income	Taste	Food Type	Rating	Parking	Waiting	Label
User_1	Restaurant_1	NJ	$500	Sweet	French	4.8	Yes	30min	1
User_1	Restaurant_2	NJ	$500	Sweet	Chinese	4.5	Yes	15min	1
User_2	Restaurant_3	NY	$600	Spicy	Mexico	4.5	Yes	20min	0
User_3	Restaurant_4	PA	$400	Salty	Fast food	3.8	No	5min	0

User Features   Item Features   Prediction

System designers: Difficult to know which feature(s) caused unfairness
Users: Difficult to know how to intervene unfair results

[3] Y Ge, J Tan, Y Zhu, Y Xia, J Luo, S Liu, Z Fu, S Geng, Z Li, and **Y Zhang**. "Explainable fairness in recommendation." In SIGIR 2022.

An Example of Yelp Recommendation

- Exposure Fairness as an Example

$$\frac{Exposure(G_0|R_{U,K})}{Exposure(G_1|R_{U,K})} = \frac{|G_0|}{|G_1|} \qquad Exposure(G_i|R_{U,K}) = \sum_{u\in U}\sum_{v\in R_{u,K}} I_{v\in G_i}$$

- Top-5 features that lead to exposure unfairness

Method	Feature-based Explanations
Pop-User	food, service, chicken, prices, hour
Pop-Item	food, service, prices, visit, hour
EFM-User	store, patio, dishes, dish, rice
EFM-Item	flavor, decor, dishes, inside, cheese
SV	server, size, pizza, food, restaurant
CEF	meal, cheese, dish, chicken, taste

[3] Y Ge, J Tan, Y Zhu, Y Xia, J Luo, S Liu, Z Fu, S Geng, Z Li, and **Y Zhang**. "Explainable fairness in recommendation." In SIGIR 2022.

RUTGERS

Counterfactual Explainable Fairness

- Explanation as a Feature Mask Vector $\Delta =$

Service	Price	Hour
1	0	0

- Simple and Effective Explanations

$$\begin{aligned}&\min.\ \text{Explanation Complexity}\\ &s.t.,\text{Model Unfairness } \leq \delta\end{aligned} \quad \Rightarrow \quad \begin{aligned}&\min.\ \|\Delta\|_1\\ &s.t.,\Psi \leq \delta\end{aligned}$$

- Ψ can be any fairness definition
  - Exposure fairness as an example

$$\frac{Exposure(G_0|R_{U,K})}{Exposure(G_1|R_{U,K})} = \frac{|G_0|}{|G_1|} \doteq \alpha \quad \Leftrightarrow \quad Exposure(G_0|R_{U,K}) = \alpha \cdot Exposure(G_1|R_{U,K})$$

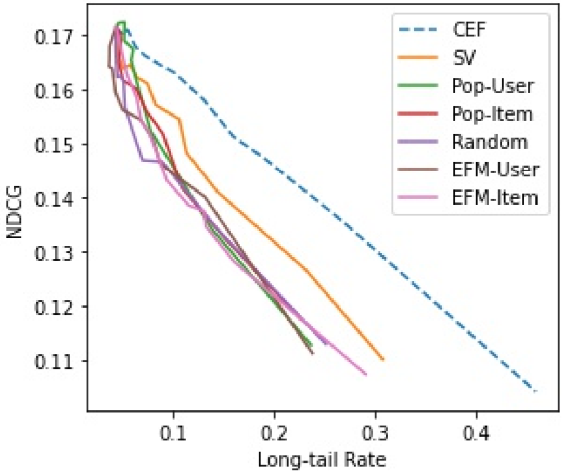
$$\min_{\Delta}.\ \|\Delta\|_1 + \lambda|\Psi|$$

$$\text{where: } \Psi = Exposure(G_0|R_{U,K}) - \alpha \cdot Exposure(G_1|R_{U,K})$$

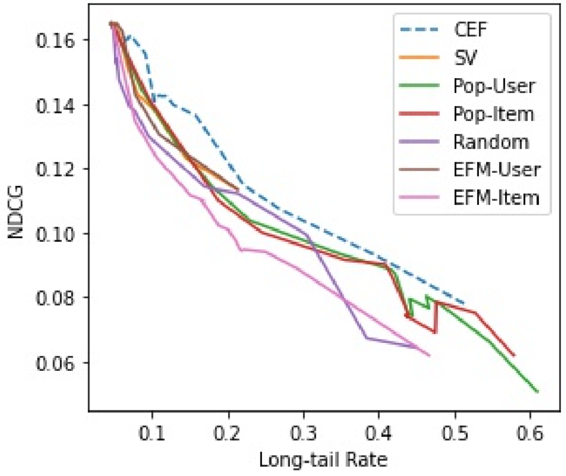
[3] Y Ge, J Tan, Y Zhu, Y Xia, J Luo, S Liu, Z Fu, S Geng, Z Li, and **Y Zhang**. "Explainable fairness in recommendation." In SIGIR 2022.

RUTGERS

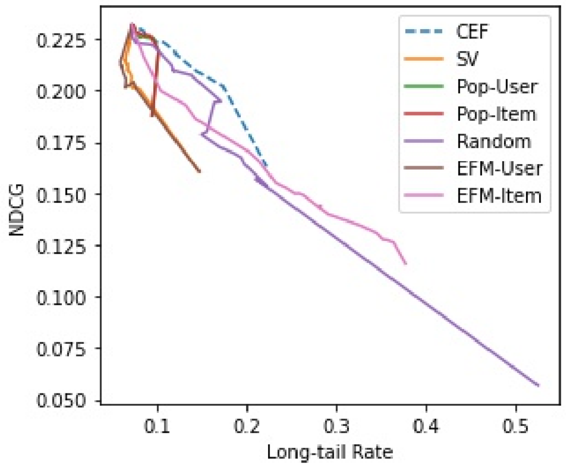
Better Fairness-Utility Trade-off



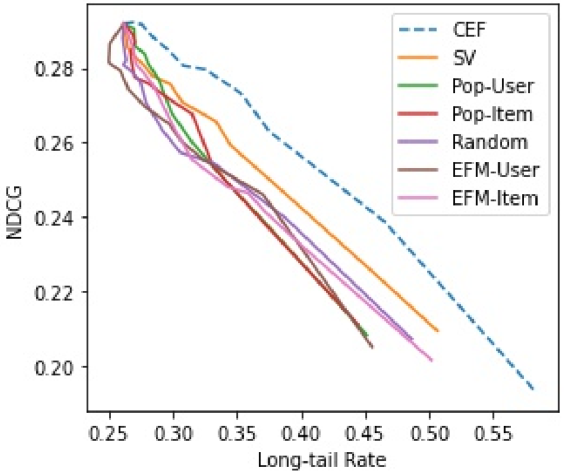
(a) NDCG@5 vs Long-tail Rate@5 on Yelp



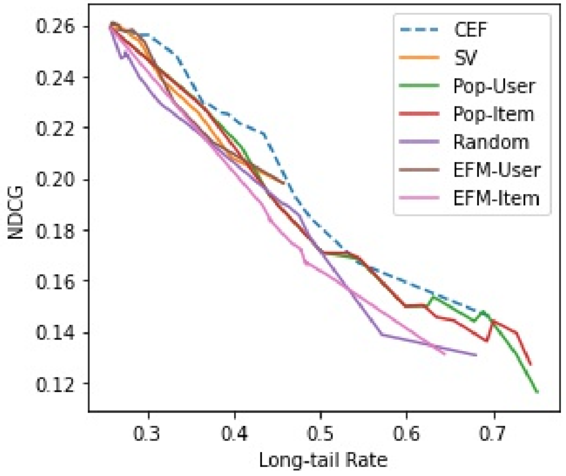
(b) NDCG@5 vs Long-tail Rate@5 on Electronics



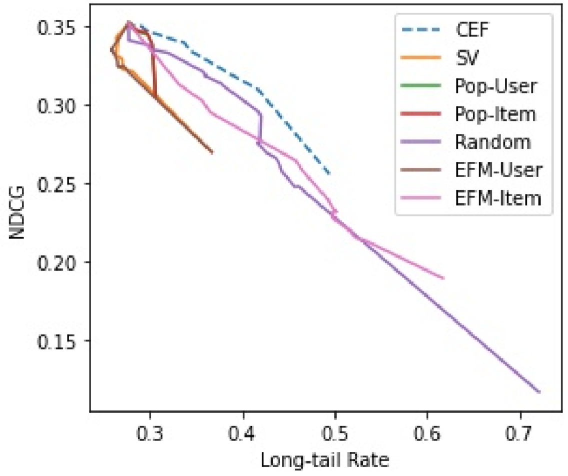
(c) NDCG@5 vs Long-tail Rate@5 on CDs&Vinyl



(d) NDCG@20 vs Long-tail Rate@20 on Yelp



(e) NDCG@20 vs Long-tail Rate@20 on Electronics



(f) NDCG@20 vs Long-tail Rate@20 on CDs&Vinyl

Trustworthy AI for Science

(ICML22, WWW22, KDD23)

[4] Z Li, J Ji, and **Y Zhang**. "From Kepler to Newton: Explainable AI for Science Discovery." In ICML AI for Science. 2022.
[5] J Tan, S Geng, Z Fu, Y Ge, S Xu, Y Li, and **Y Zhang**. "Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning." In WWW 2022.
[6] J Tan and **Y Zhang**. "ExplainableFold: Understanding AlphaFold Prediction with Explainable AI." In KDD 2023.

Science is not only about understanding the "what" and "how", but also, and perhaps more importantly, the *"why"*.

RUTGERS

The Conquest of “Why” in Science

- The conquest of why has always been the key theme of science in human history
- A Legend Example
  - The Kepler’s Laws of Planetary Motion
  - The Newton’s Law of Universal Gravitation

Tycho Brahe (1546-1610) Johannes Kepler (1571-1630) Isaac Newton (1643-1727)

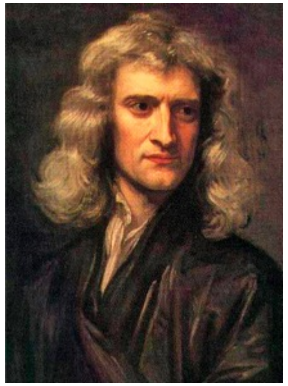
[4] Z Li, J Ji, and **Y Zhang**. "From Kepler to Newton: Explainable AI for Science Discovery." In ICML AI for Science. 2022.

RUTGERS

Three Key Roles in the Scientific Discovery Process

Tycho Brahe (1546-1610)

Johannes Kepler (1571-1630)



Isaac Newton (1643-1727)

Observation	Analyzation	Explanation

Time	Position
1	(a,b)
2	(c,d)
3	(e,f)

$$\frac{\tau^2}{r^3} = K$$

$$F = G\frac{m_1 m_2}{r^2}$$

Data Collection
Almost automated

Model Learning
Many available methods

Model Interpretation (XAI)
Still needs much exploration

[4] Z Li, J Ji, and **Y Zhang**. "From Kepler to Newton: Explainable AI for Science Discovery." In ICML AI for Science. 2022.

RUTGERS

Explainable Graph Neural Networks (WWW'22)

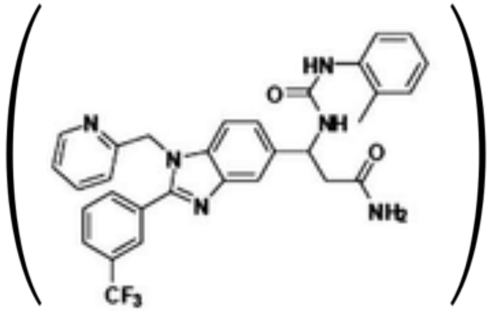
Molecule Analysis

[5] J Tan, S Geng, Z Fu, Y Ge, S Xu, Y Li, and **Y Zhang**. "Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning." In WWW 2022.

The Molecule Classification Problem

- Predict the property of molecules
  - E.g., If a molecule is soluble, toxic, or can pass the Blood-Brain Barrier
  - A fundamental problem in many tasks, e.g., drug discovery
- Molecule is a graph
  - Current approaches use Graph Neural Networks (GNN) for prediction
  - E.g., A binary classification problem

GNN

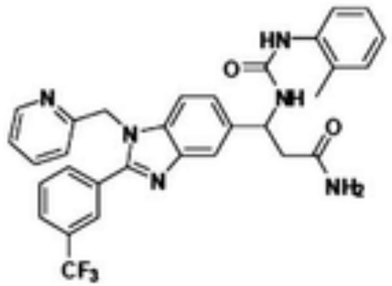


= Yes / No

- However, we want to know **why** the model produce such results

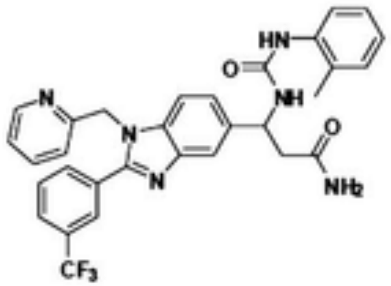
RUTGERS

Explainable Graph Neural Networks

GNN 

= Yes / No

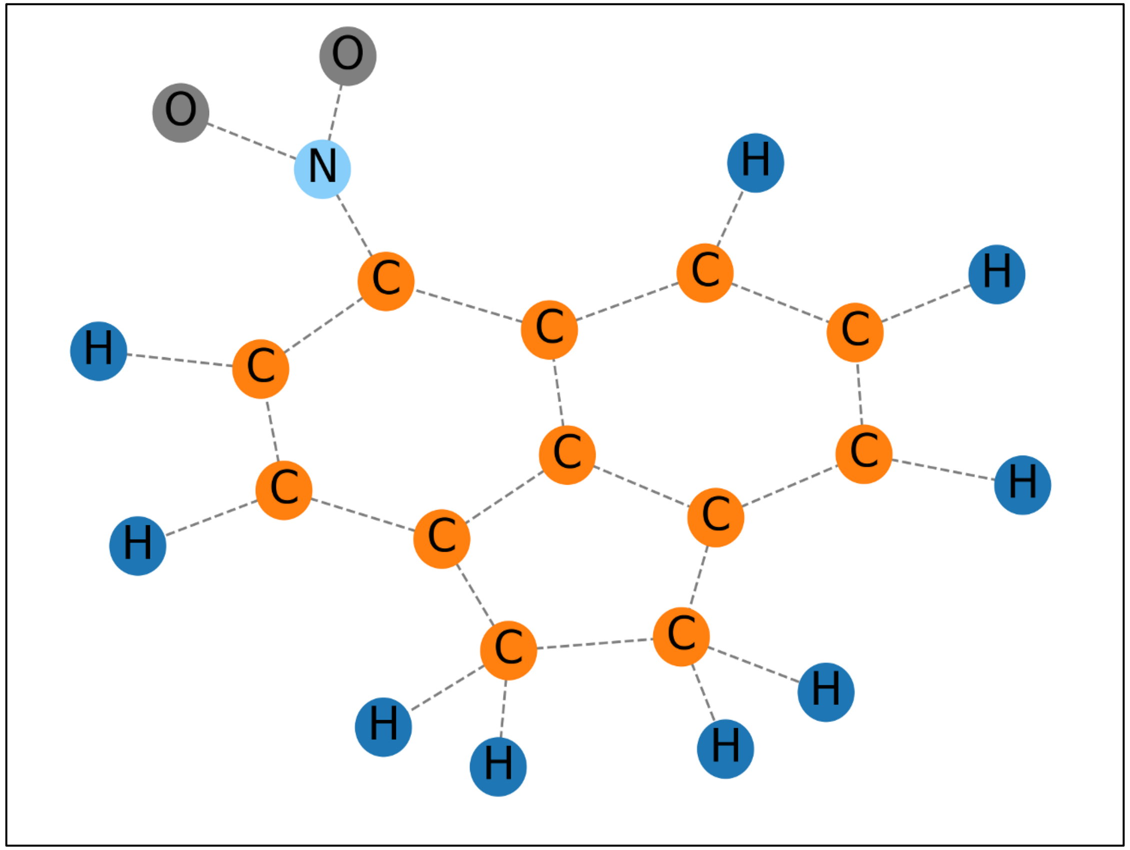
⇩

XGNN 

= Yes / No + Explanation

Factual and Counterfactual Explanations

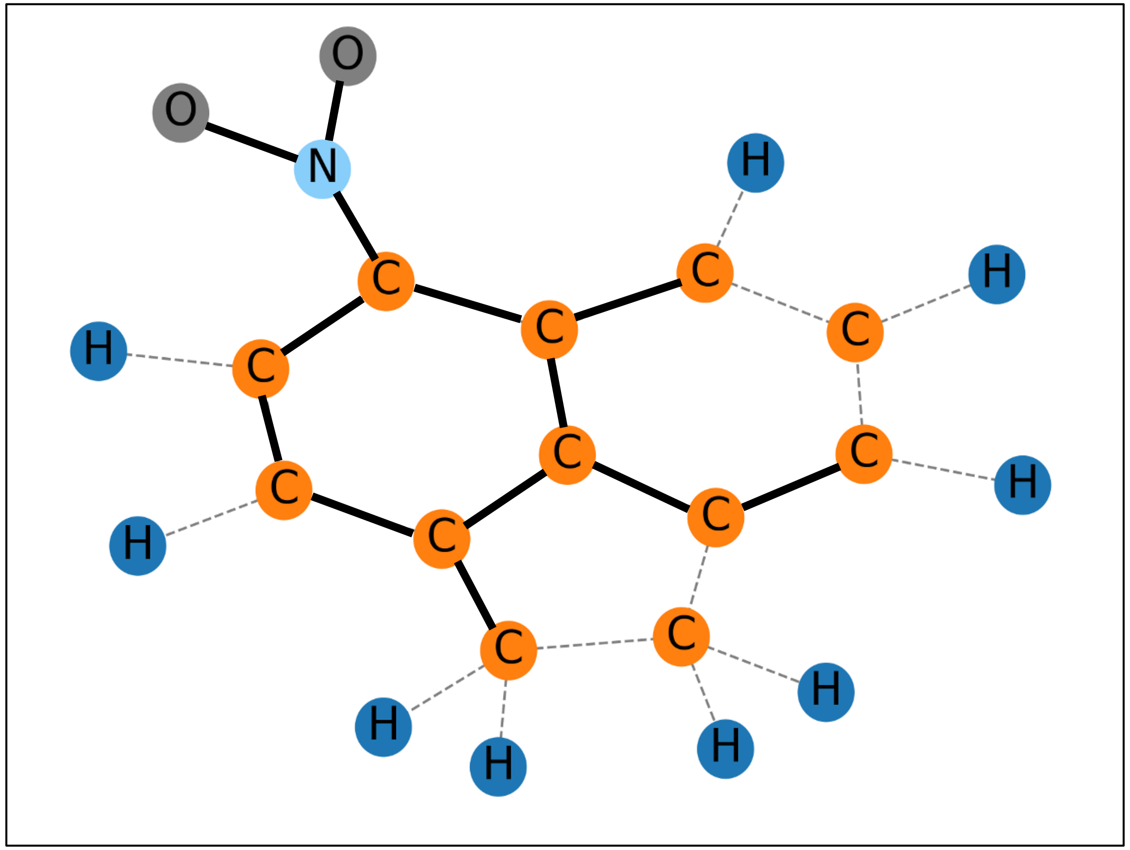
- Example: Molecule toxicity (mutagenetic) prediction [2]
  - If the GNN model predicts the molecule as toxic, why?



[5] J Tan, S Geng, Z Fu, Y Ge, S Xu, Y Li, and **Y Zhang**. "Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning." In WWW 2022.

Factual and Counterfactual Explanations

- Factual explanation seeks a sufficient condition
  - The molecule would be toxic with the highlighted bonds

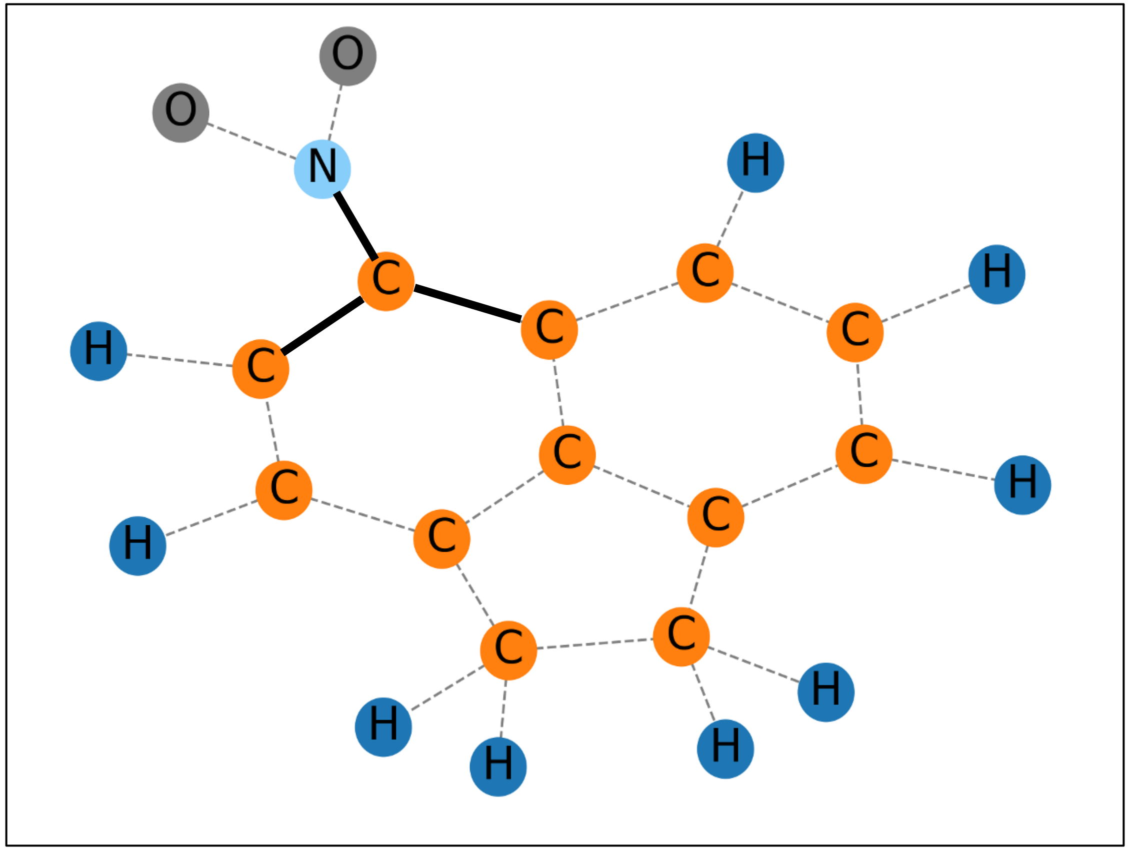


[5] J Tan, S Geng, Z Fu, Y Ge, S Xu, Y Li, and **Y Zhang**. "Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning." In WWW 2022.

RUTGERS

Factual and Counterfactual Explanations

- Counterfactual explanation seeks a necessary condition
  - The molecule would not be toxic without the highlighted bonds



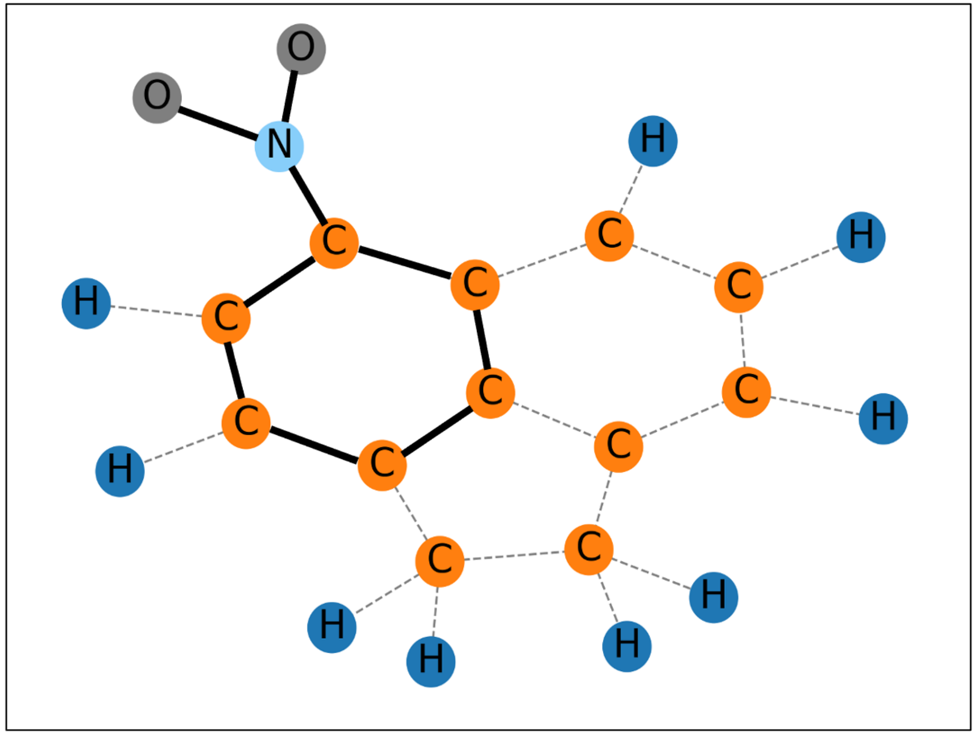
[5] J Tan, S Geng, Z Fu, Y Ge, S Xu, Y Li, and **Y Zhang**. "Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning." In WWW 2022.

Rutgers

Factual and Counterfactual Explanations

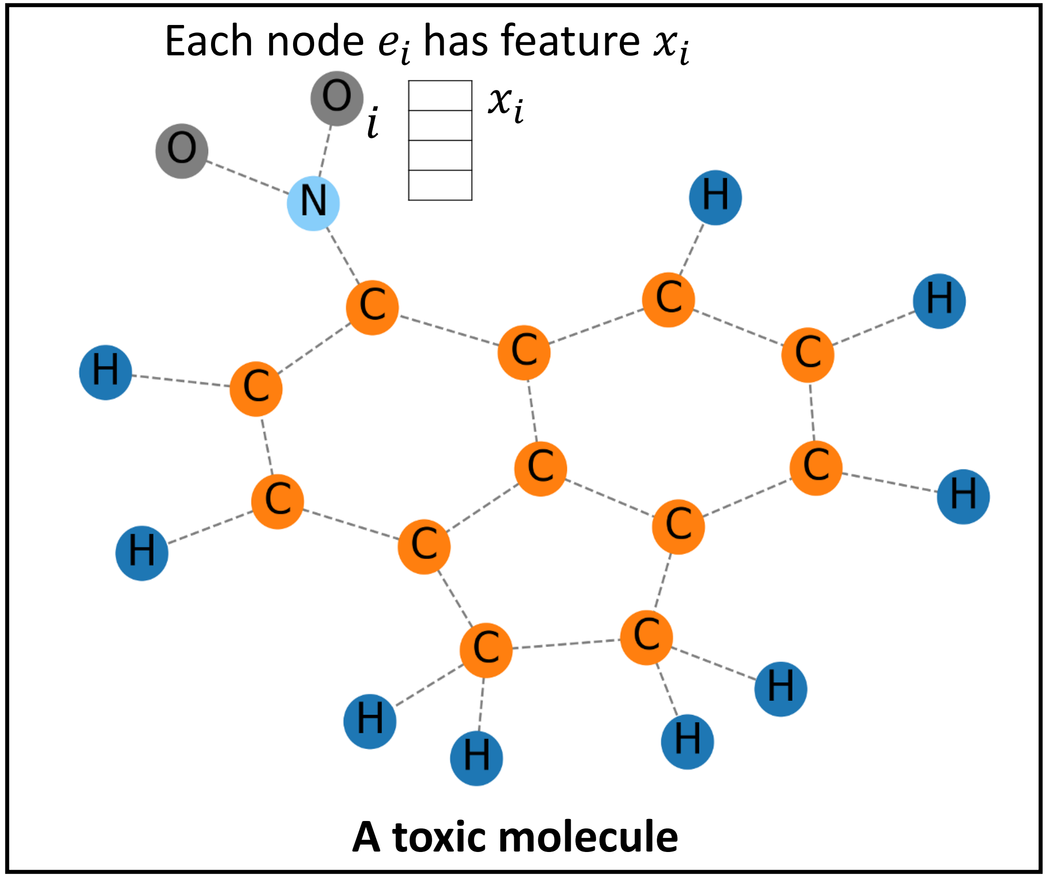
- Factual and Counterfactual explanation seeks a compact (both sufficient and necessary) condition
  - The molecule would be toxic with the highlighted bonds
  - The molecule would not be toxic without the highlighted bonds
  - No more, no less, just enough

The Nitro-Benzene Structure



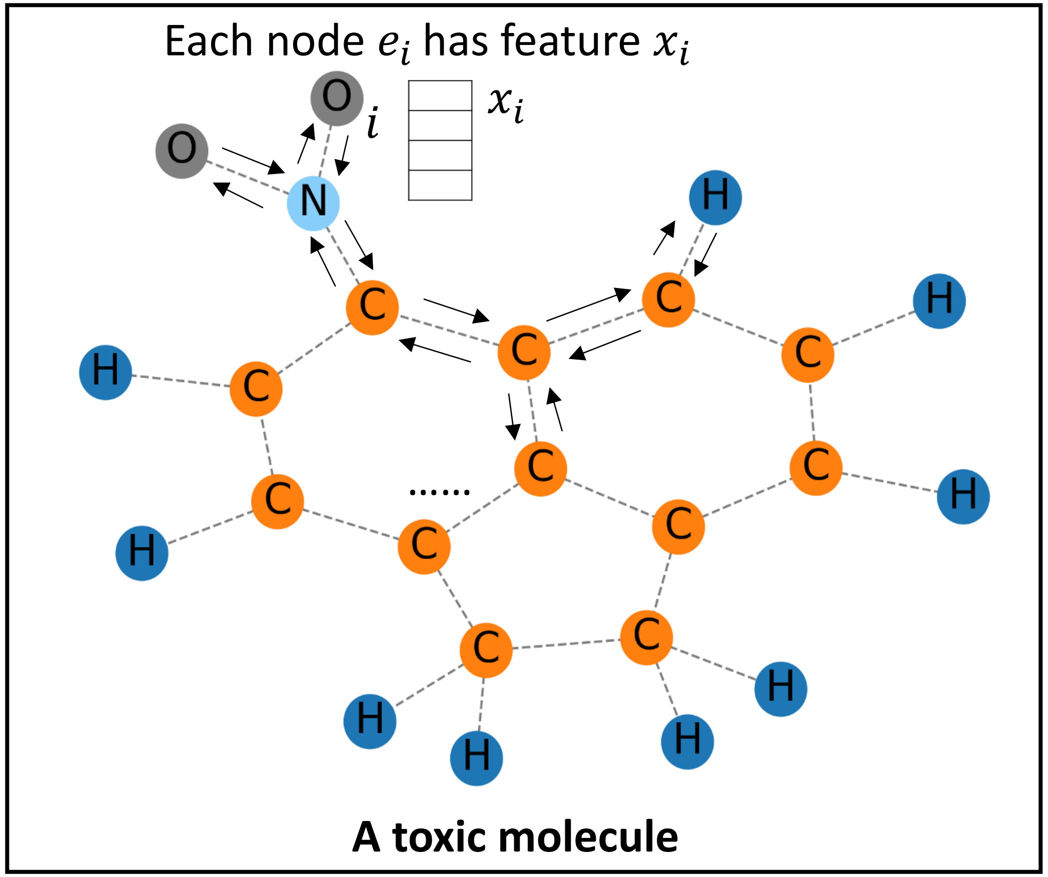
[5] J Tan, S Geng, Z Fu, Y Ge, S Xu, Y Li, and **Y Zhang**. "Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning." In WWW 2022.

GNN Basics



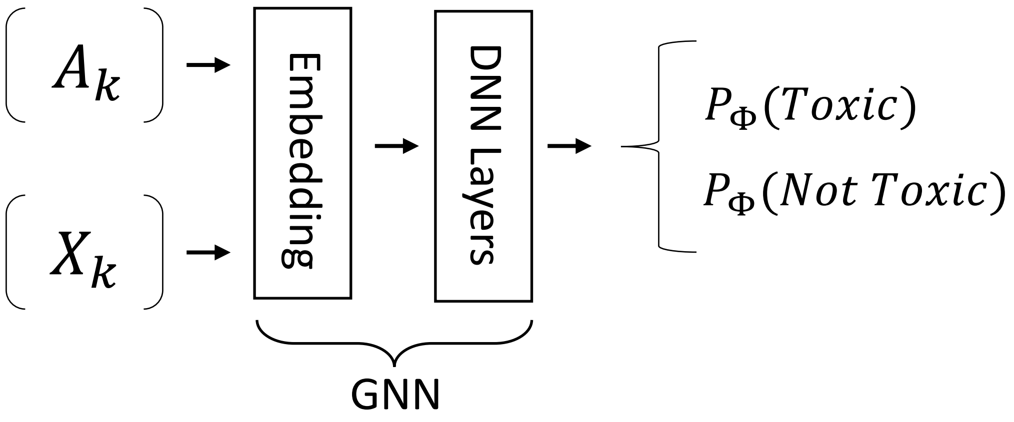
- A graph $G_k = \{\mathcal{V}_k, \mathcal{E}_k\}$
  - Adjacency matrix $A_k \in \{0,1\}^{|\mathcal{V}_k| \times |\mathcal{V}_k|}$
  - Node feature matrix $X_k \in \mathbb{R}^{|\mathcal{V}_k| \times d}$

GNN Basics



Information propagate through the graph to get graph embedding

- A graph $G_k = \{\mathcal{V}_k, \mathcal{E}_k\}$
  - Adjacency matrix $A_k \in \{0,1\}^{|\mathcal{V}_k| \times |\mathcal{V}_k|}$
  - Node feature matrix $X_k \in \mathbb{R}^{|\mathcal{V}_k| \times d}$



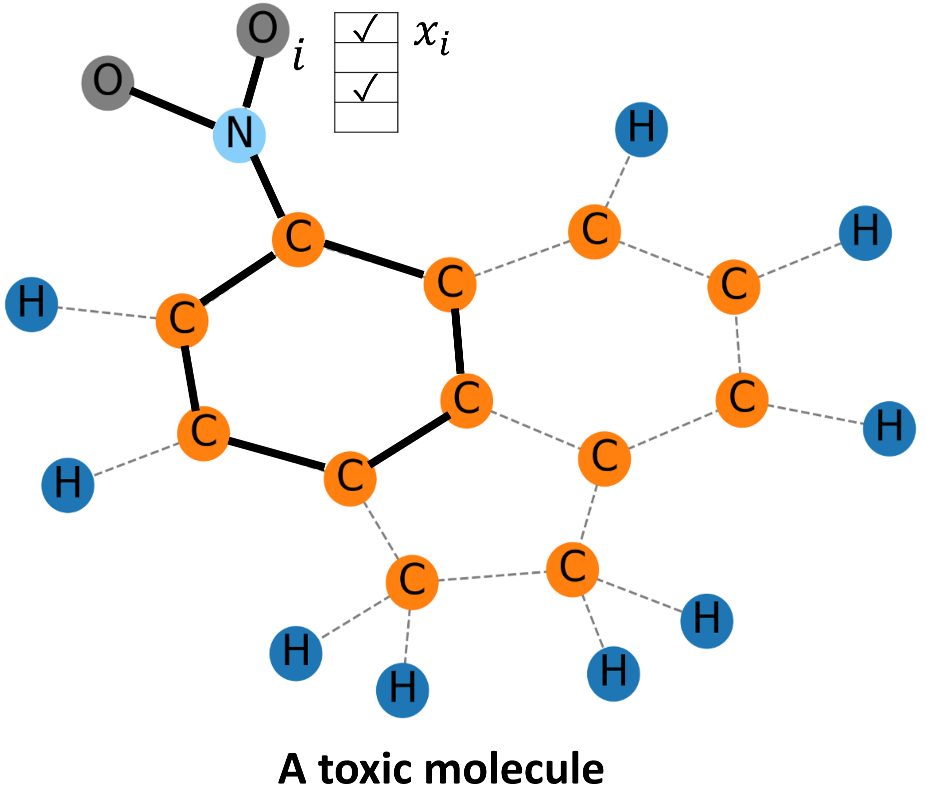
- GNN predicts the label $\hat{y}_k$ for G_k by:

$$\hat{y}_k = \arg\max_{c \in C} P_\Phi(c \mid A_k, X_k)$$

RUTGERS

GNN Explanation as Sub-Graph Mask Vector

Each node e_i has feature x_i



Explanation Sub-Graph

- A graph $G_k = \{\mathcal{V}_k, \mathcal{E}_k\}$
  - Adjacency matrix $A_k \in \{0,1\}^{|\mathcal{V}_k| \times |\mathcal{V}_k|}$
  - Node feature matrix $X_k \in \mathbb{R}^{|\mathcal{V}_k| \times d}$
- Edge mask $M_k \in \{0, 1\}^{|\mathcal{V}_k| \times |\mathcal{V}_k|}$
- Feature mask $F_k \in \{0, 1\}^{|\mathcal{V}_k| \times d}$
- Sub-Graph as Explanation
  - Sub-Edges $A_k \odot M_k$
  - Sub-Features $X_k \odot F_k$

RUTGERS

How to Find the Explanation?

- Factual Reasoning: Given A already happened, will B happen?
  - Factual Condition:

$$\arg\max_{c \in C} P_{\Phi}(c \mid \underbrace{A_k \odot M_k, X_k \odot F_k}_{\text{With only the explanation sub-graph}}) = \hat{y}_k$$

- Counterfactual Reasoning: If A did not happen, would B still happen?
  - Counterfactual Condition:

$$\arg\max_{c \in C} P_{\Phi}(c \mid \underbrace{A_k - A_k \odot M_k, X_k - X_k \odot F_k}_{\text{Without the explanation sub-graph}}) \neq \hat{y}_k$$

RUTGERS

What are Good Explanations? Simple and Effective (again!)

Occam's Razor Principle for Explainable AI:
When trying to explain a phenomenon, if two explanations are equally effective, then we prefer the simpler one.

- To quantify Simplicity
  - Explanation Complexity

$$C(M, F) = \|M\|_0 + \|F\|_0$$

How many **edges** are included in the explanation

How many **features** are included in the explanation

- To quantify Effectiveness
  - Factual Explanation Strength

$$S_f(M, F) = P_\Phi(\hat{y}_k \mid A_k \odot M_k, X_k \odot F_k)$$

  - Counterfactual Explanation Strength

$$S_c(M, F) = -P_\Phi(\hat{y}_k \mid A_k - A_k \odot M_k, X_k - X_k \odot F_k)$$

Both should be large enough to satisfy the conditions

RUTGERS

Counterfactual Learning and Reasoning

- Seek simple and effective explanations

$$\text{minimize} \;\; \text{Explanation Complexity}$$
$$\text{s.t.,} \;\; \text{Explanation is Strong Enough}$$

⇩

$$\text{minimize} \;\; C(M_k, F_k)$$
$$\text{s.t.,} \;\; S_f(M_k, F_k) > P_\Phi(\hat{y}_{k,s} \mid A_k \odot M_k, X_k \odot F_k),$$
$$S_c(M_k, F_k) > -P_\Phi(\hat{y}_{k,s} \mid A_k - A_k \odot M_k, X_k - X_k \odot F_k)$$

- $\hat{y}_{k,s}$ is the label of the second largest prediction probability
- Idea: Find minimal components of a molecule which is both sufficient and necessary

Evaluation of Counterfactual Explanations

Sufficiency and Necessity:

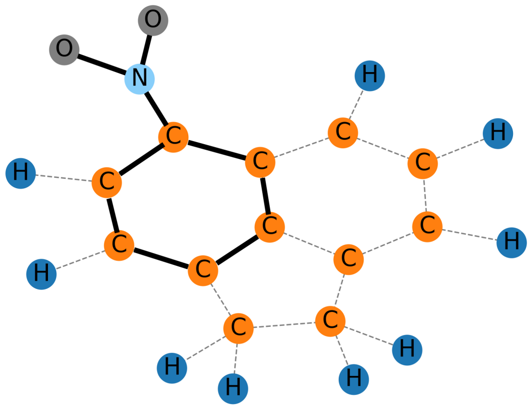
S ⇒ N: S is a sufficient condition for N

¬N ⇒ ¬S: N is a necessary condition for S

- Probability of Sufficient (PS)
  - If we only keep the explanation sub-graph, the prediction result is the same, then the explanation is sufficient
  - PS: Percentage of molecules whose explanation sub-graph is Sufficient

$$\mathrm{PS} = \frac{\sum_{G_k \in \mathcal{G}} \mathrm{ps}_k}{|\mathcal{G}|}, \text{ where } \mathrm{ps}_k = \begin{cases} 1, & \text{if } \hat{y}'_k = \hat{y}_k \\ 0, & \text{else} \end{cases}$$

$$\text{where } \hat{y}'_k = \arg\max_{c \in C} P_\Phi(c \mid A_k \odot M_k, X_k \odot F_k)$$



RUTGERS

Evaluation of Counterfactual Explanations

Sufficiency and Necessity:

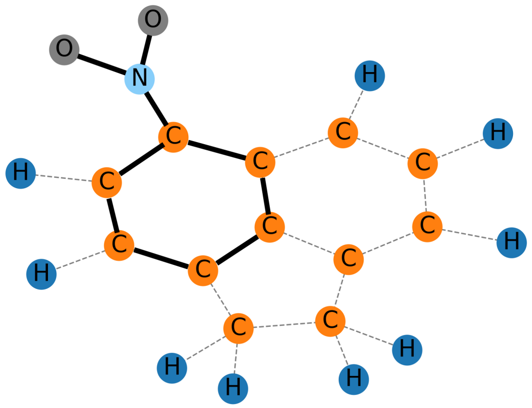
S ⇒ N: S is a sufficient condition for N

¬N ⇒ ¬S: N is a necessary condition for S

- Probability of Necessity (PN)
  - If we remove the explanation sub-graph, the prediction result will change, then the explanation is necessary
  - PN: Percentage of molecules whose explanation sub-graph is Necessary

$$\mathrm{PN} = \frac{\sum_{G_k \in \mathcal{G}} \mathrm{pn}_k}{|\mathcal{G}|}, \text{ where } \mathrm{pn}_k = \begin{cases} 1, & \text{if } \hat{y}'_k \neq \hat{y}_k \\ 0, & \text{else} \end{cases}$$

$$\text{where } \hat{y}'_k = \arg\max_{c \in C} P_\Phi(c \mid A_k - A_k \odot M_k, X_k - X_k \odot F_k)$$

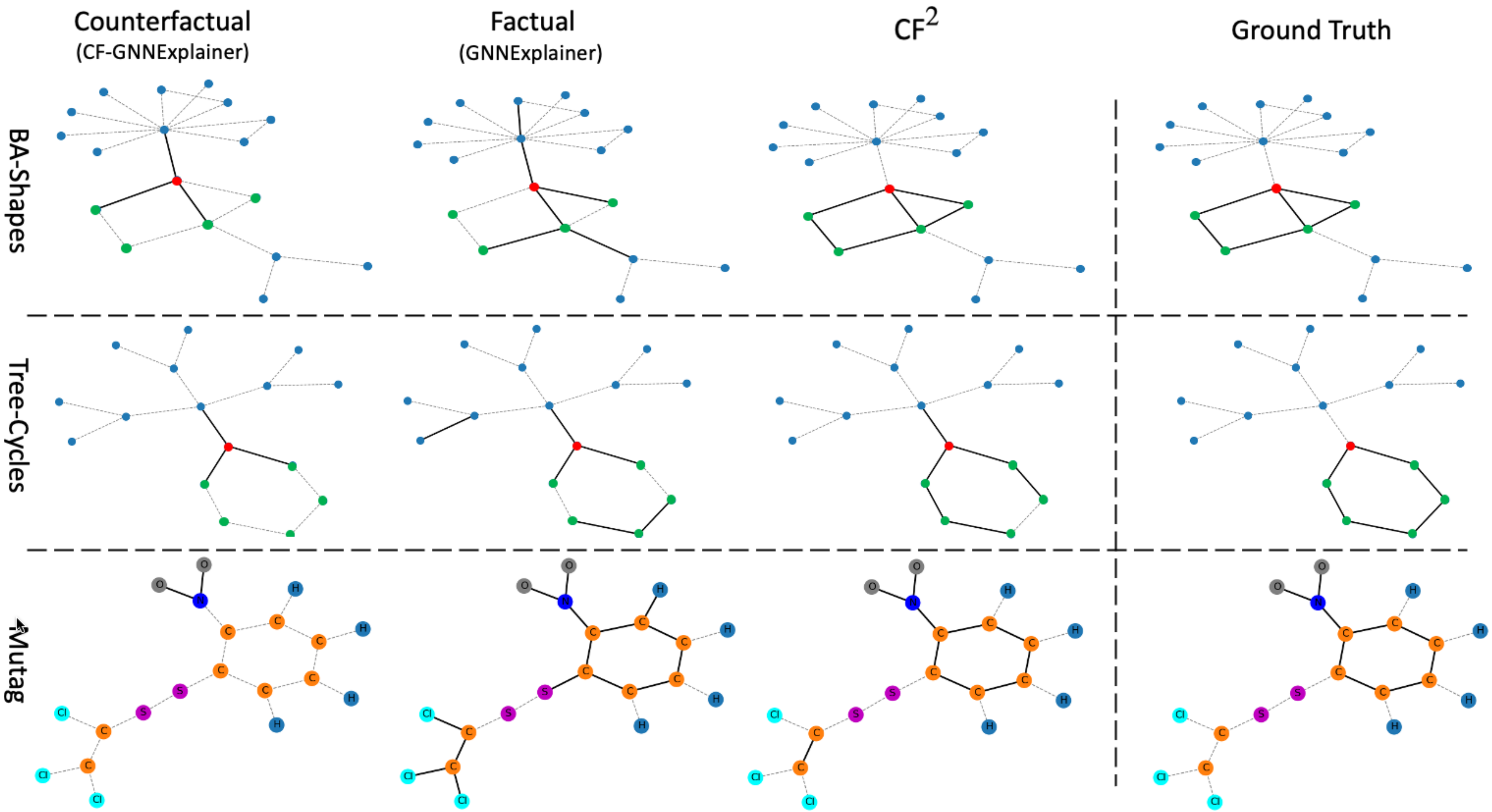


Datasets for Evaluation

Dataset	#graph	#ave n	#ave e	#class	#feat	task	gt
BA-Shapes	1	700	4100	4	-	node	✓
Tree-Cycles	1	871	1950	2	-	node	✓
Mutag	4337	30.32	30.77	2	14	graph	
$Mutag_0$	2301	31.74	32.54	2	14	graph	✓
NCI1	4110	29.87	32.30	2	37	graph	
CiteSeer	1	3312	4732	6	3703	node	

RUTGERS

Qualitative Case Study



Evaluation with PN, PS

- This evaluation does not need ground-truth explanation

Models	BA-Shapes				Tree-Cycles				Mutag$_0$			
	PN%	PS%	F_{NS}%	#exp	PN%	PS%	F_{NS}%	#exp	PN%	PS%	F_{NS}%	#exp
GNNExplainer†	72.19	45.62	55.91	6.00	100.00	59.72	74.78	6.00	71.79	97.44	82.67	15.00
CF-GNNExplainer	75.34	41.10	53.18	5.79	100.00	31.94	48.42	3.44	96.26	7.48	13.88	7.72
Gem†	61.36	52.27	56.45	6.00	100.00	29.89	46.02	6.00	83.01	76.42	79.58	15.00
CF^2	76.73	68.22	**72.07**	6.21	100.00	81.94	**90.08**	5.81	97.44	100.00	**98.70**	14.95

Models	NCI1				CiteSeer (edge)				CiteSeer (feature)			
	PN%	PS%	F_{NS}%	#exp	PN%	PS%	F_{NS}%	#exp	PN%	PS%	F_{NS}%	#exp
GNNExplainer†	92.13	62.16	74.24	15.00	66.67	90.05	76.61	5.00	71.64	99.50	72.79	60.00
CF-GNNExplainer	97.14	31.43	47.49	7.75	69.50	82.00	75.23	2.58	72.14	92.54	81.07	72.91
Gem†	99.03	52.15	68.32	15.00	61.05	72.67	66.36	5.00	-	-	-	-
CF^2	100.00	63.81	**77.91**	17.70	71.00	94.50	**81.08**	3.18	74.63	95.02	**83.60**	62.73

Evaluate with Accuracy

- This evaluation needs ground-truth explanation

Models	BA-Shapes				Tree-Cycles				Mutag_0			
	Acc%	Pr%	Re%	F_1%	Acc%	Pr%	Re%	F_1%	Acc%	Pr%	Re%	F_1%
GNNExplainer†	95.25	60.08	60.08	60.08	92.78	68.06	68.06	68.06	96.96	59.71	85.17	68.85
CF-GNNExplainer	94.39	67.19	54.11	56.79	90.27	87.40	47.45	59.10	96.91	66.09	39.46	47.39
Gem†	**96.97**	64.16	64.16	64.16	89.88	57.23	57.23	57.23	96.43	63.12	47.11	54.68
CF^2	96.37	73.15	68.18	**66.61**	**93.26**	84.92	73.84	**75.69**	**97.34**	65.28	88.59	**72.56**

Kendall's τ and Spearman's ρ correlation between Accuracy and PN, PS

Models	BA-Shapes		Tree-Cycles		Mutag_0	
	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$
F_{NS} & F_1	1.00	1.00	1.00	1.00	1.00	1.00
F_{NS} & Acc	0.66	0.79	1.00	1.00	0.66	0.79

$$Pr = \frac{TP}{TP+FP} \quad Re = \frac{TP}{TP+FN} \quad Acc = \frac{TP+TN}{ALL}$$

$$F_1 = \frac{2Pr \cdot Re}{Pr + Re} \quad F_{NS} = \frac{2PN \cdot PS}{PN + PS}$$

PN/PS-based evaluation is highly correlated with ground-truth-based evaluation.

We can use PN/PS to evaluate explanations when ground-truth is not available

Rutgers

ExplainableFold (KDD'23)

Understanding AlphaFold Prediction with Explainable AI

[6] J Tan and **Y Zhang**. "ExplainableFold: Understanding AlphaFold Prediction with Explainable AI." In KDD 2023.

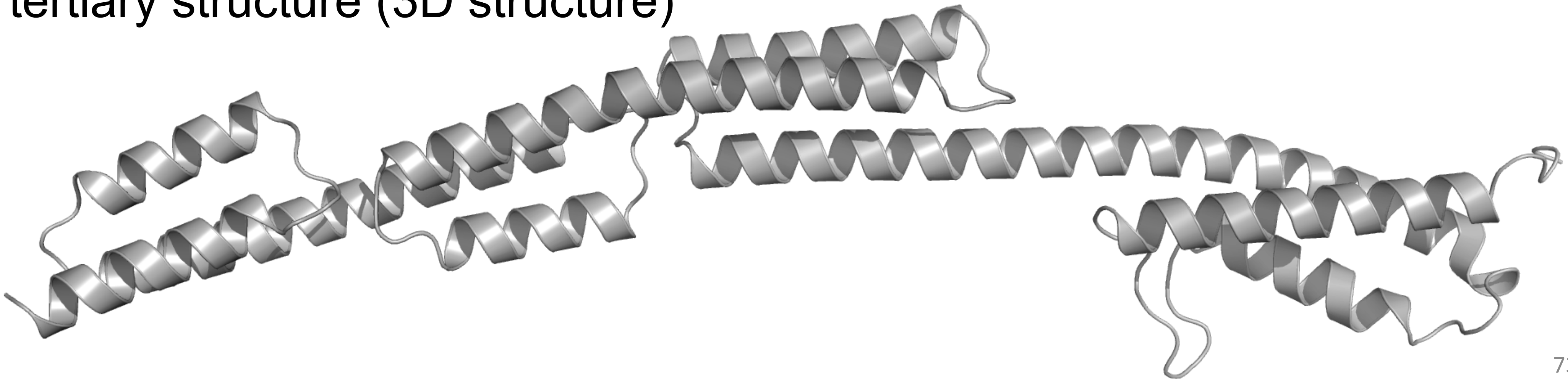
The Protein Folding Problem

From primary structure (amino acid sequence)

D Q E L G K Q S R R S Q D I I K S L …… D L T R G T K E D

Cryo-Electron Microscopy
(Cryo-EM)

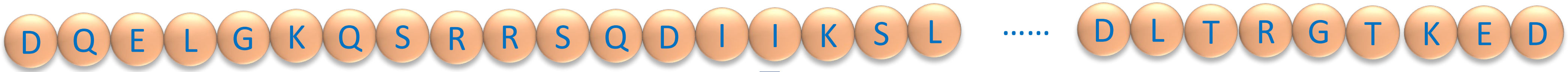
To tertiary structure (3D structure)

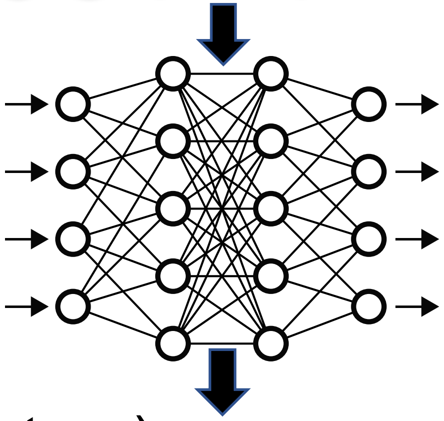


RUTGERS

The Protein Folding Problem

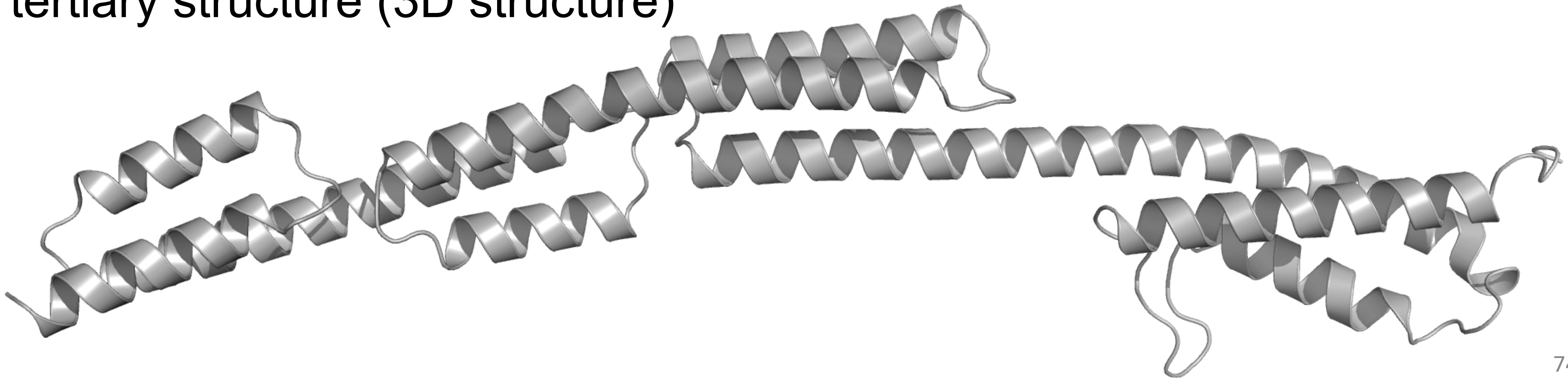
From primary structure (amino acid sequence)





AlphaFold revolutionizes Protein Structure Prediction

To tertiary structure (3D structure)

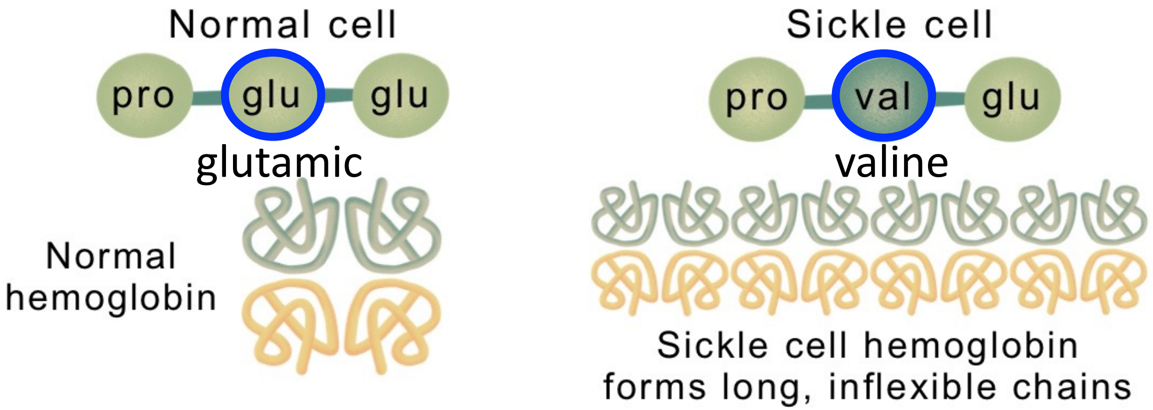


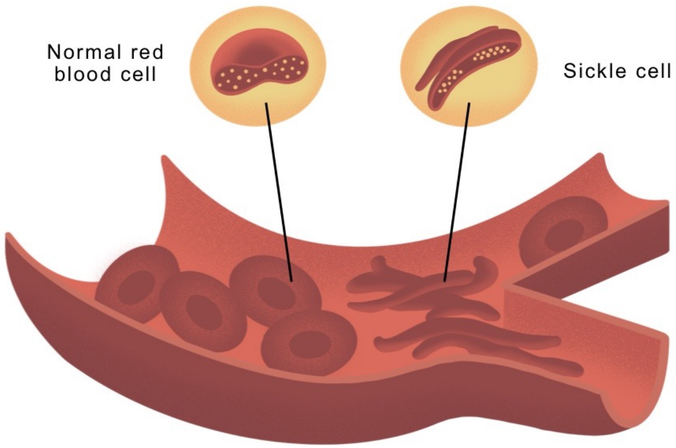
RUTGERS

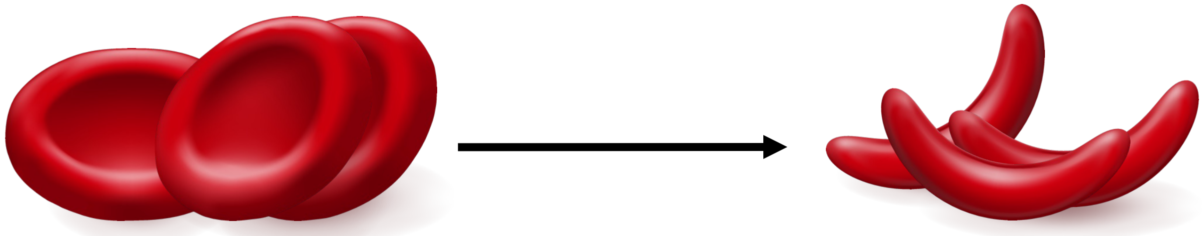
Science is not
only about understanding the “what” and “how”,
but also, and perhaps more importantly, the
“why”.

Explanation Provides Important Insights for Scientists

- Cause-effect explanation between amino-acid and protein structure
  - One single substitution in the HBB gene can significantly change the structure of hemoglobin, causing the sickle-cell anemia







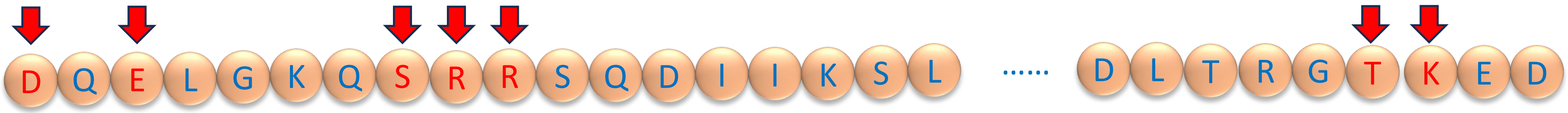
Certain amino acids play significant roles in the protein folding process!

[5] Rees, David C., Thomas N. Williams, and Mark T. Gladwin. "Sickle-cell disease." The Lancet 376.9757 (2010): 2018-2031. https://sickle-cell.com/

RUTGERS

ExplainableFold Problem Definition

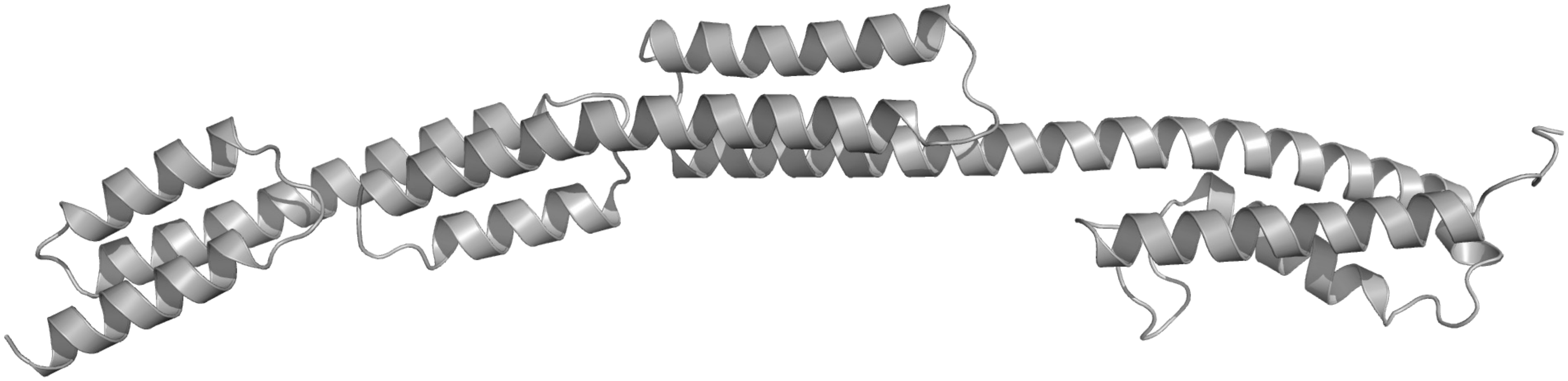
Identify the most crucial residues that cause the proteins to fold into the structures they are [6].



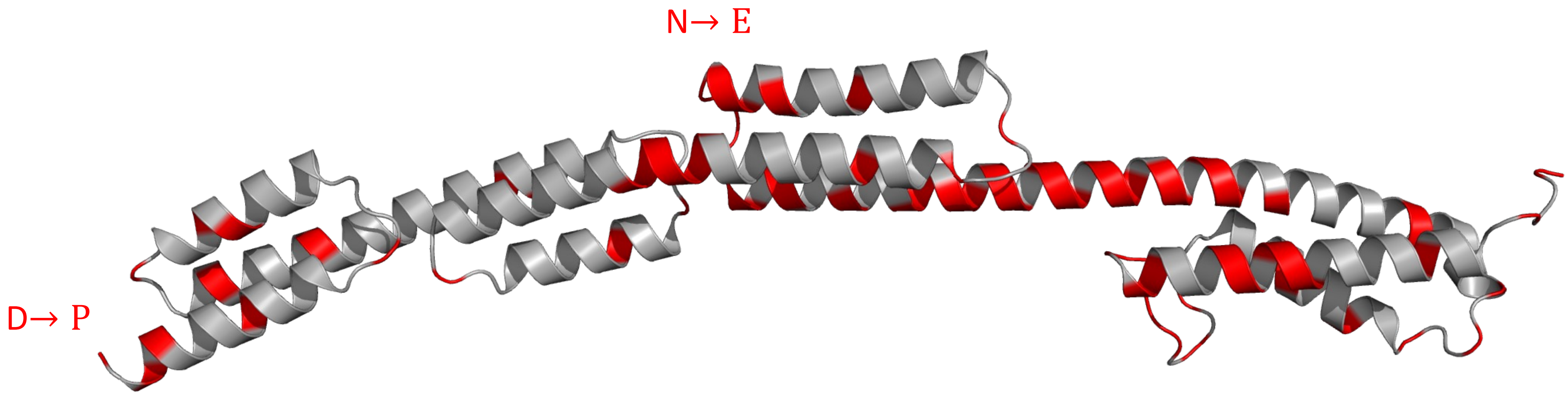
[6] J Tan and **Y Zhang**. "ExplainableFold: Understanding AlphaFold Prediction with Explainable AI." In KDD 2023.

RUTGERS

ExplainableFold Problem Definition

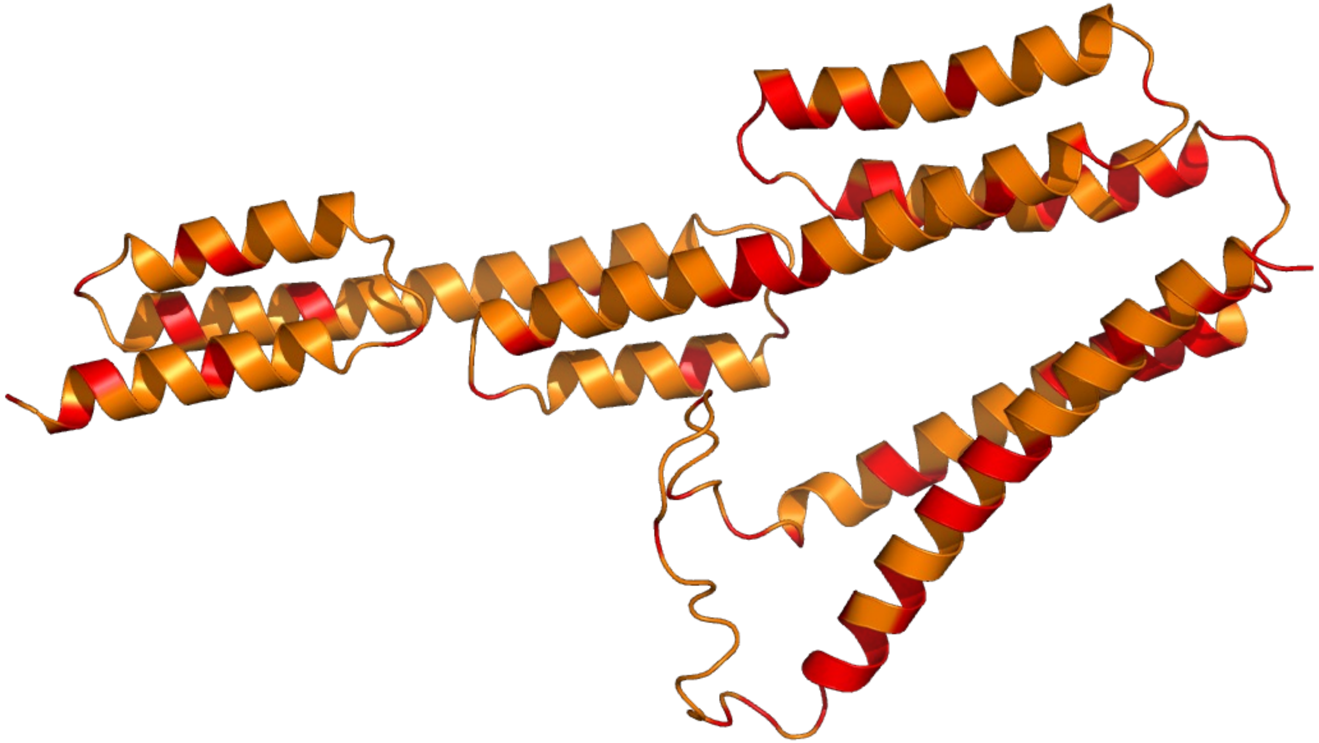


ExplainableFold Problem Definition



RUTGERS

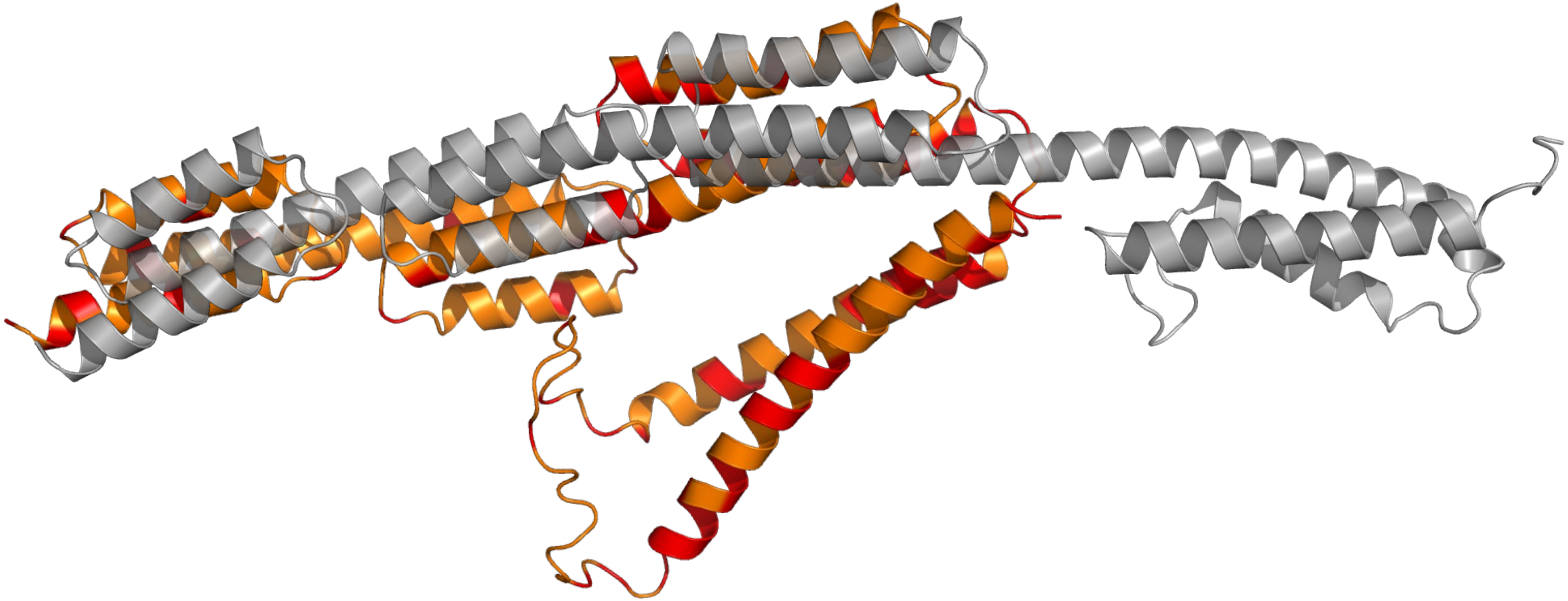
ExplainableFold Problem Definition



ExplainableFold Problem Definition

TM-score: 0.44 (TM<0.5 means different folding structure [7,8])

TM-score = Template Modeling score



[7] Jinrui Xu and Yang Zhang. How significant is a protein structure similarity with tm-score= 0.5? Bioinformatics, 26(7):889–895, 2010.
[8] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. Proteins: Structure, Function, and Bioinformatics, 57(4):702–710, 2004.

RUTGERS

What are Good Explanations? Simple and Effective (again!)

Occam's Razor Principle for Explainable AI:
When trying to explain a phenomenon, if two explanations are equally effective, then we prefer the simpler one.

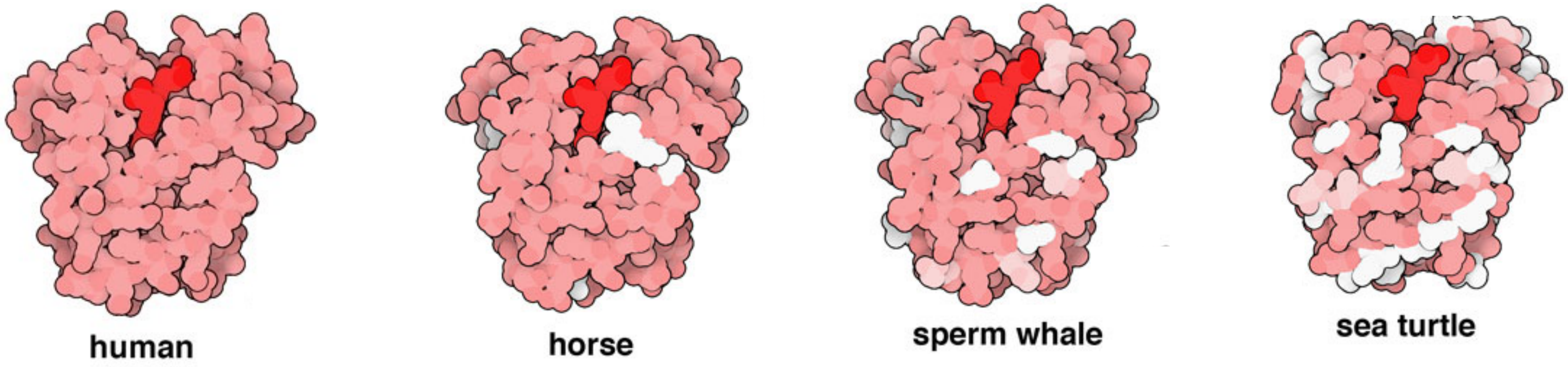
- For a target protein $\boldsymbol{P}$, $\boldsymbol{P} \in \{0,1\}^{21\times l}$, MSA $\boldsymbol{M}(\boldsymbol{P}) \in \{0,1\}^{m\times 21\times l}$
- We learn a counterfactual protein embedding $\boldsymbol{P}'$

$$\begin{aligned} &\text{minimize} \;\; \text{Explanation Complexity} \\ &\text{s.t.,} \;\; \text{Explanation is Strong Enough} \end{aligned} \Rightarrow \begin{aligned} &\text{minimize } ||\boldsymbol{P} - \boldsymbol{P}'||_0 && \text{Simple} \\ &\text{s.t. } \mathrm{TM}(\boldsymbol{S}, \boldsymbol{S}') \le 0.5, \;\; \boldsymbol{P}' \in \{0,1\}^{21\times l} && \text{Effective} \\ &\text{where } \boldsymbol{S}' = f_\theta\big(\boldsymbol{P}', \boldsymbol{M}(\boldsymbol{P}')\big) && \text{Blackbox (AlphaFold)} \end{aligned}$$

Evaluation

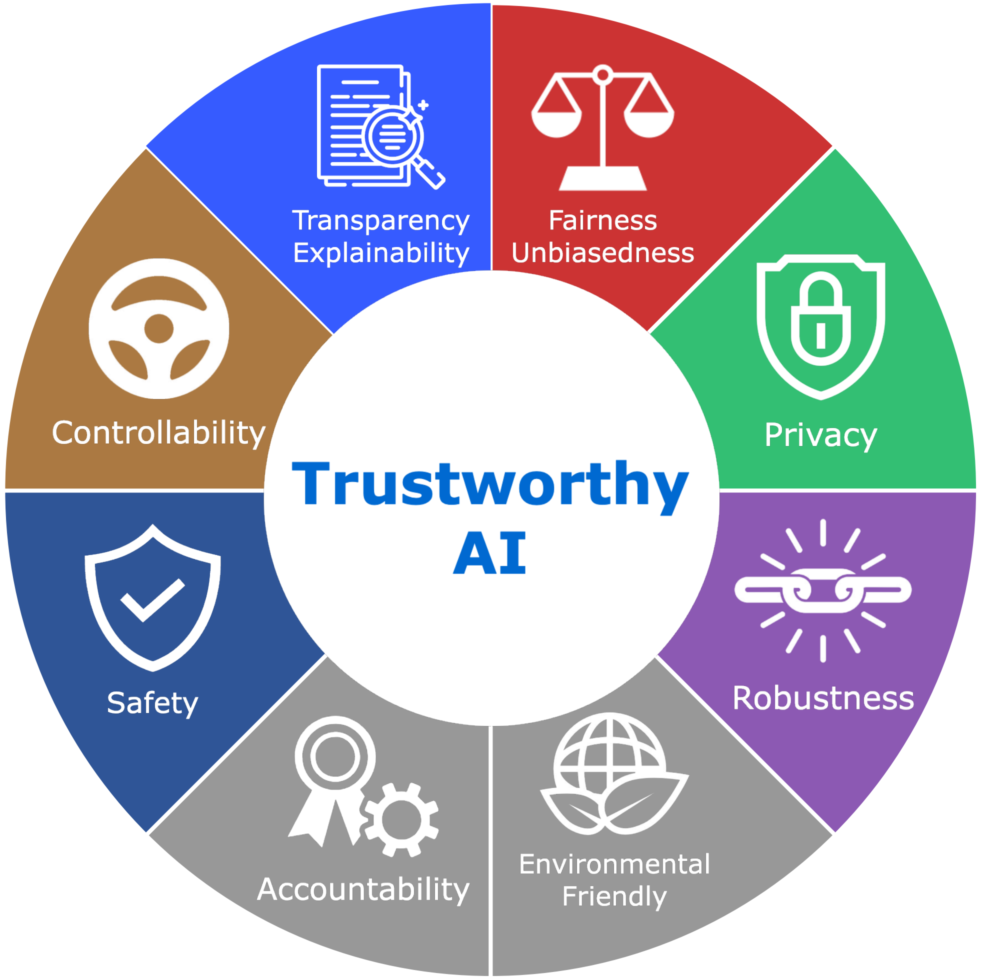
- CASP-14 Dataset (same as AlphaFold): 152 target proteins

	Ave Explanation Size ($\lvert\mathcal{E}\rvert$) ↓	Ave Complexity ($\lvert\mathcal{E}\rvert/l$) ↓	Ave TM-score $TM(S, S^*)$ ↓	PN score↑
Random	85.22	0.33	0.83	0.07
Evolutionary [40]	88.42	0.33	0.77	0.16
ExplainableFold	**83.33**	**0.31**	**0.59**	**0.40**

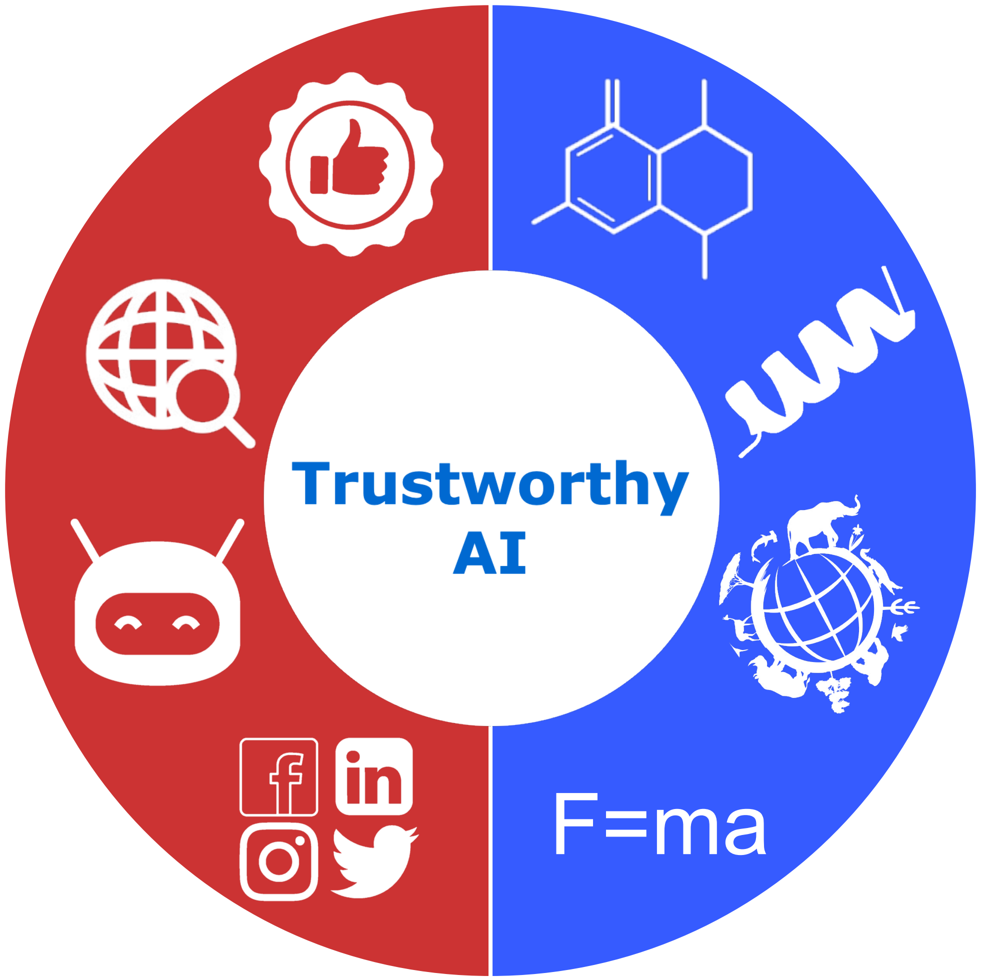


[9] 14th Critical Assessment of protein Structure Prediction. Moult, J., K. Fidelis, A. Kryshtafovych, T. Schwede, and Maya Topf. "Critical assessment of techniques for protein structure prediction, fourteenth round." CASP 14 Abstract Book (2020).

Summary



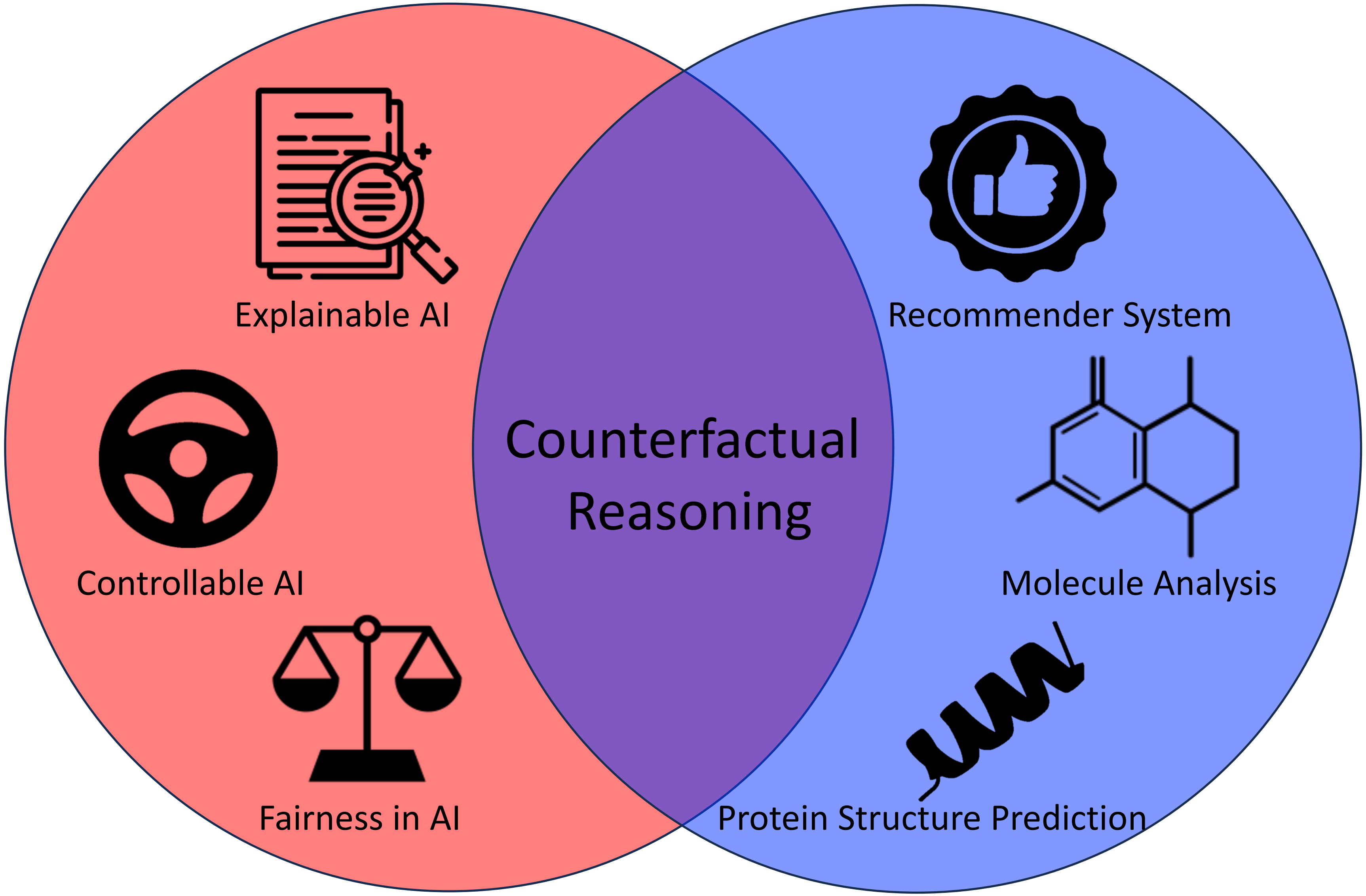
Methodology



Applications

RUTGERS

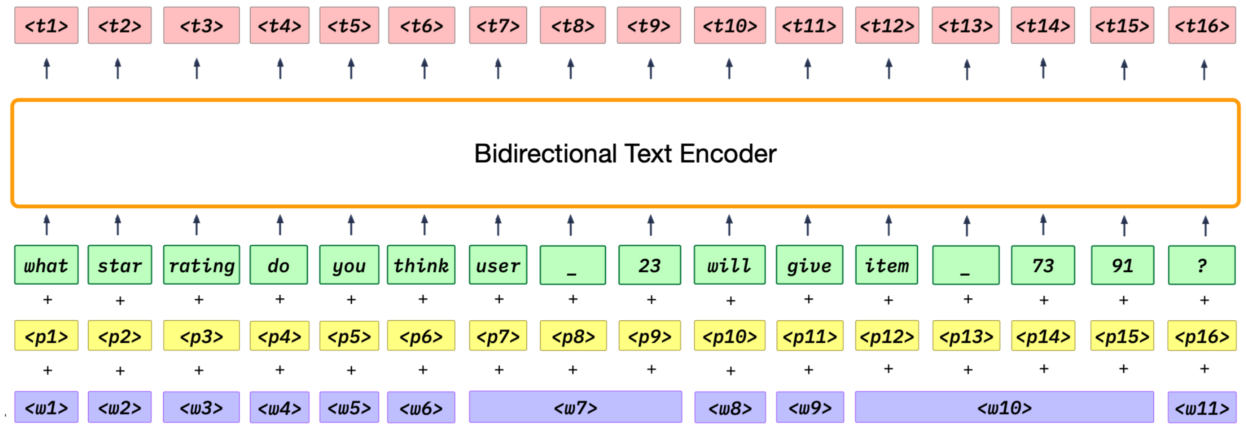
Summary



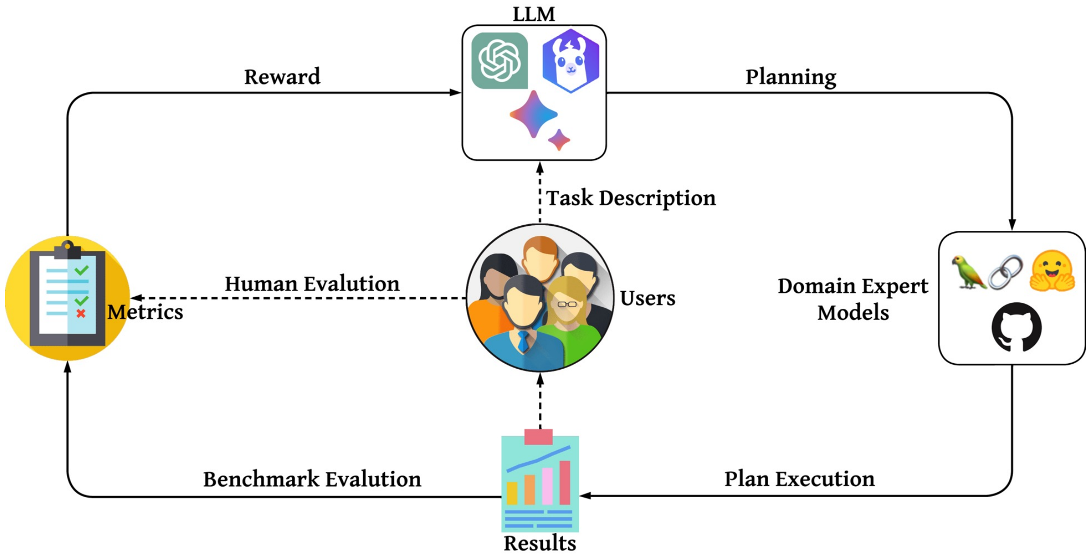
Multiple Trustworthy Perspectives **Multiple Application Domains**

RUTGERS

Future Research



Trustworthy Large Language Models (LLMs) [9]



OpenAGI: Trustworthy Autonomous AI Agents [10]

[9] W Hua, Y Ge, S Xu, J Ji, and **Y Zhang**. "UP5: Unbiased Foundation Model for Fairness-aware Recommendation." arXiv:2305.12090 (2023).
[10] Y Ge, W Hua, K Mei, J Ji, J Tan, S Xu, Z Li and **Y Zhang**. "OpenAGI: When LLM Meets Domain Experts." arXiv:2304.04370 (2023).

Yongfeng Zhang

Department of Computer Science, Rutgers University

yongfeng.zhang@rutgers.edu

http://yongfeng.me

Reference

- [KDD23] Tan, Juntao, Yongfeng Zhang. "ExplainableFold: Understanding AlphaFold Prediction with Explainable AI." KDD. 2023.
- [ECAI23] Tan, Juntao, Yingqiang Ge, Yan Zhu, Yinglong Xia, Jiebo Luo, Jianchao Ji, and Yongfeng Zhang. "User-Controllable Recommendation via Counterfactual Retrospective and Prospective Explanations." ECAI. 2023.
- [WSDM23] Ji, Jianchao, Zelong Li, Shuyuan Xu, Max Xiong, Juntao Tan, Yingqiang Ge, Hao Wang, and Yongfeng Zhang. "Counterfactual Collaborative Reasoning." WSDM. 2023.
- [ICLR23] Geng, Shijie, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. "Hiclip: Contrastive language-image pretraining with hierarchy-aware attention." ICLR. 2023.
- [ICTIR23] Xu, Shuyuan, Yingqiang Ge, Yunqi Li, Zuohui Fu, Xu Chen, and Yongfeng Zhang. "Causal collaborative filtering." In Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, pp. 235-245. 2023.
- [TORS23] Xu, Shuyuan, Juntao Tan, Shelby Heinecke, Vena Jia Li, and Yongfeng Zhang. "Deconfounded causal collaborative filtering." ACM Transactions on Recommender Systems. 2023.
- [TIST23a] Li, Yunqi, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. "Fairness in Recommendation: Foundations, Methods and Applications." ACM Transactions on Intelligent Systems and Technology (2023).
- [TIST23b] Li, Lei, Yongfeng Zhang, and Li Chen. "On the relationship between explanation and recommendation: Learning to rank explanations for improved performance." ACM Transactions on Intelligent Systems and Technology 14, no. 2 (2023): 1-24.
- [SIGIR23] Chen, Ziheng, Fabrizio Silvestri, Jia Wang, Yongfeng Zhang, and Gabriele Tolomei. "The dark side of explanations: Poisoning recommender systems with counterfactual examples." SIGIR. 2023.
- [TOIS23a] Xu, Zhichao, Hansi Zeng, Juntao Tan, Zuohui Fu, Yongfeng Zhang, Qingyao Ai. "A Reusable Model-agnostic Framework for Faithfully Explainable Recommendation and System Scrutability." ACM Transactions on Information Systems. 2023.
- [TOIS23b] Li, Lei, Yongfeng Zhang, and Li Chen. "Personalized prompt learning for explainable recommendation." TOIS. 2023.

- [TACL23] Hua, Wenyue, Lifeng Jin, Linfeng Song, Haitao Mi, Yongfeng Zhang, and Dong Yu. "Discover, Explanation, Improvement: Automatic Slice Detection Framework for Natural Language Processing." TACL. 2023.
- [CIKM23] Li, Lei, Yongfeng Zhang, and Li Chen. "Prompt Distillation for Efficient LLM-based Recommendation." CIKM. 2023.
- [WWW22a] Tan, Juntao, Shijie Geng, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Yunqi Li, and Yongfeng Zhang. "Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning." WWW. 2022.
- [WWW22b] Geng, Shijie, Zuohui Fu, Juntao Tan, Yingqiang Ge, Gerard De Melo, and Yongfeng Zhang. "Path language modeling over knowledge graphs for explainable recommendation." In Proceedings of the ACM Web Conference 2022, pp. 946-955. 2022.
- [WWW22c] Wen, Bingbing, Yunhe Feng, Yongfeng Zhang, and Chirag Shah. "ExpScore: Learning metrics for recommendation explanation." In Proceedings of the ACM Web Conference 2022, pp. 3740-3744. 2022.
- [ICML22] Li, Zelong, Jianchao Ji, and Yongfeng Zhang. "From Kepler to Newton: Explainable AI for Science Discovery." In ICML 2022 2nd AI for Science Workshop. 2022.
- [SIGIR22a] Ge, Yingqiang, Juntao Tan, Yan Zhu, Yinglong Xia, Jiebo Luo, Shuchang Liu, Zuohui Fu, Shijie Geng, Zelong Li, and Yongfeng Zhang. "Explainable fairness in recommendation." SIGIR. 2022.
- [SIGIR22b] Li, Zelong, Jianchao Ji, Yingqiang Ge, and Yongfeng Zhang. "AutoLossGen: Automatic loss function generation for recommender systems." SIGIR. 2022.
- [CIKM22a] Xu, Shuyuan, Juntao Tan, Zuohui Fu, Jianchao Ji, Shelby Heinecke, Yongfeng Zhang. "Dynamic causal collaborative filtering." In Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022.
- [CIKM22b] Chen, Hanxiong, Yunqi Li, He Zhu, and Yongfeng Zhang. "Learn basic skills and reuse: Modularized adaptive neural architecture search (manas)." CIKM. 2022.
- [RecSys22a] Geng, Shijie, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. "Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5)." RecSys. 2022.

- [RecSys22b] Liu, Shuchang, Yingqiang Ge, Shuyuan Xu, Yongfeng Zhang, and Amelie Marian. "Fairness-aware federated matrix factorization." In Proceedings of the 16th ACM Conference on Recommender Systems, pp. 168-178. 2022.
- [ACL22] Geng, Shijie, Zuohui Fu, Yingqiang Ge, Lei Li, Gerard De Melo, and Yongfeng Zhang. "Improving personalized explanation generation through visualization." ACL. 2022.
- [ICML22] Li, Zelong, Jianchao Ji, and Yongfeng Zhang. "From Kepler to Newton: Explainable AI for Science Discovery." In ICML 2022 2nd AI for Science Workshop. 2022.
- [EMNLP22a] Hua, Wenyue, and Yongfeng Zhang. "System 1+ System 2= Better World: Neural-Symbolic Chain of Logic Reasoning." In Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 601-612. 2022.
- [EMNLP22b] Fang, Yanbo, and Yongfeng Zhang. "Data-Efficient Concept Extraction from Pre-trained Language Models for Commonsense Explanation Generation." In Findings of the Association for Computational Linguistics: EMNLP 2022. 2022.
- [AACL22] Fang, Yanbo, Zuohui Fu, Xin Luna Dong, Yongfeng Zhang, Gerard de Melo. "Assessing Combinational Generalization of Language Models in Biased Scenarios." In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing. 2022.
- [JCDL22] Li, Yunqi, Hanxiong Chen, Juntao Tan, and Yongfeng Zhang. "Causal factorization machine for robust recommendation." In Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries, pp. 1-9. 2022.
- [WSDM22a] Chen, Hanxiong, Yunqi Li, Shaoyun Shi, Shuchang Liu, He Zhu, and Yongfeng Zhang. "Graph collaborative reasoning." In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, pp. 75-84. 2022.
- [WSDM22b] Ge, Yingqiang, Xiaoting Zhao, Lucia Yu, Saurabh Paul, Diane Hu, Chu-Cheng Hsieh, and Yongfeng Zhang. "Toward Pareto efficient fairness-utility trade-off in recommendation through reinforcement learning." WSDM. 2022.
- [WSDM22c] Ovaisi, Zohreh, Shelby Heinecke, Jia Li, Yongfeng Zhang, Elena Zheleva, and Caiming Xiong. "RGRecSys: A toolkit for robustness evaluation of recommender systems." WSDM. 2022.

- [WWW21a] Chen, Hanxiong, Shaoyun Shi, Yunqi Li, and Yongfeng Zhang. "Neural collaborative reasoning." In Proceedings of the Web Conference 2021, pp. 1516-1527. 2021.
- [WWW21b] Li, Yunqi, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. "User-oriented fairness in recommendation." In Proceedings of the Web Conference 2021, pp. 624-632. 2021.
- [ACL21] Li, Lei, Yongfeng Zhang, and Li Chen. "Personalized Transformer for Explainable Recommendation." In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 4947-4957. 2021.
- [CIKM21a] Tan, Juntao, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang. "Counterfactual explainable recommendation." In Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021.
- [CIKM21b] Fu, Zuohui, Yikun Xian, Shijie Geng, Gerard De Melo, and Yongfeng Zhang. "Popcorn: Human-in-the-loop popularity debiasing in conversational recommender systems." CIKM. 2021.
- [RecSys21] Xian, Yikun, Tong Zhao, Jin Li, Jim Chan, Andrey Kan, Jun Ma, Xin Luna Dong, Christos Faloutsos, George Karypis, S. Muthukrishnan, Yongfeng Zhang. "Ex3: Explainable attribute-aware item-set recommendations." RecSys. 2021.
- [SIGIR21a] Li, Yunqi, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. "Towards personalized fairness based on causal notion." In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1054-1063. 2021.
- [SIGIR21b] Liu, Shuchang, Shuyuan Xu, Wenhui Yu, Zuohui Fu, Yongfeng Zhang, and Amelie Marian. "FedCT: Federated collaborative transfer for recommendation." SIGIR. 2021.
- [SIGIR21c] Fu, Zuohui, Yikun Xian, Yaxin Zhu, Shuyuan Xu, Zelong Li, Gerard De Melo, and Yongfeng Zhang. "Hoops: Human-in-the-loop graph reasoning for conversational recommendation." SIGIR. 2021.
- [SIGIR21d] Li, Lei, Yongfeng Zhang, and Li Chen. "Extra: Explanation ranking datasets for explainable recommendation." SIGIR. 2021.

- [WSDM21] Ge, Yingqiang, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei et al. "Towards long-term fairness in recommendation." WSDM. 2021.
- [NAACL21] Zhu, Yaxin, Yikun Xian, Zuohui Fu, Gerard de Melo, and Yongfeng Zhang. "Faithfully Explainable Recommendation via Neural Logic Reasoning." In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021.
- [CIKM20a] Shi, Shaoyun, Hanxiong Chen, Weizhi Ma, Jiaxin Mao, Min Zhang, and Yongfeng Zhang. "Neural logic reasoning." In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 1365-1374. 2020.
- [CIKM20b] Xian, Yikun, Zuohui Fu, Handong Zhao, Yingqiang Ge, Xu Chen, Qiaoying Huang, Shijie Geng, Zhou Qin, Gerard de Melo, S. Muthukrishnan, Yongfeng Zhang. "CAFE: Coarse-to-fine neural symbolic reasoning for explainable recommendation." CIKM. 2020.
- [CIKM20c] Li, Lei, Yongfeng Zhang, and Li Chen. "Generate neural template explanations for recommendation." In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 755-764. 2020.
- [COLING20] Mukadam, Meet, Mandhara Jayaram, and Yongfeng Zhang. "A Representation Learning Approach to Animal Biodiversity Conservation." In Proceedings of the 28th International Conference on Computational Linguistics, pp. 294-305. 2020.
- [SIGIR20a] Fu, Zuohui, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, Gerard de Melo. "Fairness-aware explainable recommendation over knowledge graphs." SIGIR. 2020.
- [SIGIR20b] Ge, Yingqiang, Shuya Zhao, Honglu Zhou, Changhua Pei, Fei Sun, Wenwu Ou, and Yongfeng Zhang. "Understanding echo chambers in e-commerce recommender systems." SIGIR. 2020.
- [SIGIR20c] Ge, Yingqiang, Shuyuan Xu, Shuchang Liu, Zuohui Fu, Fei Sun, and Yongfeng Zhang. "Learning personalized risk preferences for recommendation." In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 409-418. 2020.

- [WWW20] Li, Lei, Li Chen, and Yongfeng Zhang. "Towards controllable explanation generation for recommender systems via neural template." In Companion proceedings of the web conference 2020, pp. 198-202. 2020.
- [UbiComp20] Wang, Guang, Yongfeng Zhang, Zhihan Fang, Shuai Wang, Fan Zhang, and Desheng Zhang. "FairCharge: A data-driven fairness-aware charging recommendation system for large-scale electric taxi fleets." UbiComp. 2020.
- [SIGIR19a] Xian, Yikun, Zuohui Fu, Shan Muthukrishnan, Gerard De Melo, and Yongfeng Zhang. "Reinforcement knowledge graph reasoning for explainable recommendation." SIGIR. 2019.
- [SIGIR19b] Chen, Xu, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. "Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation." SIGIR. 2019.
- [SIGIR19c] Qu, Chen, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. "BERT with history answer embedding for conversational question answering." SIGIR. 2019.
- [WWW19a] Ge, Yingqiang, Shuyuan Xu, Shuchang Liu, Shijie Geng, Zuohui Fu, and Yongfeng Zhang. "Maximizing marginal utility per dollar for economic recommendation." In The World Wide Web Conference, pp. 2757-2763. 2019.
- [WWW19b] Pei, Changhua, Xinru Yang, Qing Cui, Xiao Lin, Fei Sun, Peng Jiang, Wenwu Ou, and Yongfeng Zhang. "Value-aware recommendation based on reinforcement profit maximization." In The World Wide Web Conference, pp. 3123-3129. 2019.
- [TOIS19] Ai, Qingyao, Yongfeng Zhang, Keping Bi, and W. Bruce Croft. "Explainable product search with a dynamic relation embedding model." ACM Transactions on Information Systems (TOIS) 38, no. 1 (2019): 1-29.
- [CIKM19a] Bi, Keping, Qingyao Ai, Yongfeng Zhang, and W. Bruce Croft. "Conversational product search based on negative feedback." In Proceedings of the 28th acm international conference on information and knowledge management, 2019.

RUTGERS

- [CIKM19b] Qu, Chen, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W. Bruce Croft, and Mohit Iyyer. "Attentive history selection for conversational question answering." In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 1391-1400. 2019.
- [AAAI19] Chen, Xu, Yongfeng Zhang, and Zheng Qin. "Dynamic explainable recommendation based on neural attentive models." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 53-60. 2019.
- [CIKM18] Zhang, Yongfeng, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. "Towards conversational search and recommendation: System ask, user respond." In Proceedings of the 27th acm international conference on information and knowledge management. 2018.
- [SIGIR17] Ai, Qingyao, Yongfeng Zhang, Keping Bi, Xu Chen, and W. Bruce Croft. "Learning a hierarchical embedding model for personalized product search." In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 645-654. 2017.
- [RecSys17] Xiao, Lin, Min Zhang, Yongfeng Zhang, Zhaoquan Gu, Yiqun Liu, and Shaoping Ma. "Fairness-aware group recommendation with pareto-efficiency." In Proceedings of the eleventh ACM conference on recommender systems. 2017.
- [WSDM17] Zhao, Qi, Yongfeng Zhang, Yi Zhang, and Daniel Friedman. "Multi-product utility maximization for economic recommendation." In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, 2017.
- [WWW16] Zhang, Yongfeng, Qi Zhao, Yi Zhang, Daniel Friedman, Min Zhang, Yiqun Liu, and Shaoping Ma. "Economic recommendation with surplus maximization." In Proceedings of the 25th international conference on world wide web, 2016.
- [IJCAI15] Zhang, Yongfeng, Yunzhi Tan, Min Zhang, Yiqun Liu, Tat-Seng Chua, and Shaoping Ma. "Catch the black sheep: unified framework for shilling attack detection based on fraudulent action propagation." In Twenty-fourth international joint conference on artificial intelligence. 2015.
- [SIGIR14] Zhang, Yongfeng, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis." In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, pp. 83-92. 2014.