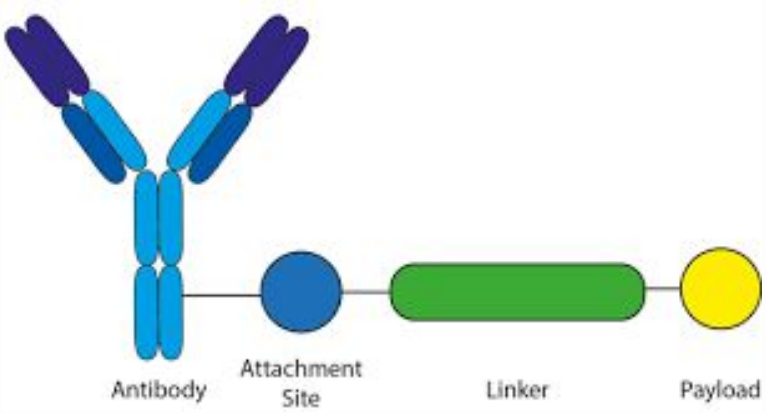
Generating Novel Linker Structures in Antibody-Drug Conjugates Using Diffusion Models

By: April Cao

Introduction

- Nearly 10 million deaths from cancer in 2022
- Antibody-drug conjugates (ADCs)
  - New form of cancer treatment in the last 30 years
  - 3 main components:
    - Monoclonal antibody (mAb)
    - Payload
    - Linker



Background

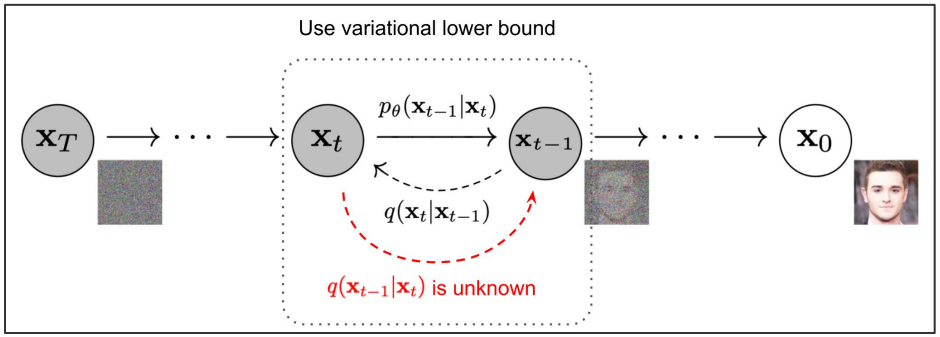
Linker

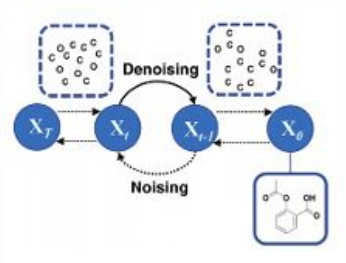
- Short chemical sequence that connects the antibody and payload
- Crucial for regulating stability and releasing payload in the tumor
    - Also affects pharmacokinetics, efficacy, and toxicity characteristics

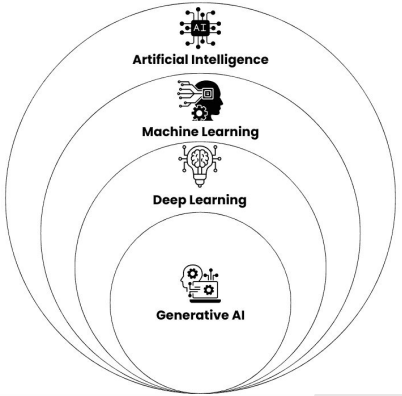
- Hurdles in linker development:
    - Complex structure and susceptibility to premature release can cause systemic toxicity
    - Number of parameters that can be investigated
        - Stability at physiological pH and in serum, ability to release payload where desired, sufficient hydrophilicity, reasonable/practical manufacturing process, etc.

Background

- **Machine Learning**: rapidly analyzing datasets to identify patterns
- **Generative AI**: model creates original data by using existing data as reference
- ★ **Diffusion Models:** generative model that starts with random noise and gradually refines its data over many steps
  - Model learns how data naturally changes over time, and then reverses that process to create realistic outputs
  - Diffusion models have been used in a variety of bioinformatics problems, eg., protein design







Aims

1. Develop a diffusion-based generative model to generate novel linker structures in ADCs
2. Validate generated linker structures and their ADMET (absorption, distribution, metabolism, excretion, toxicity) properties to evaluate the stability of the novel sequences

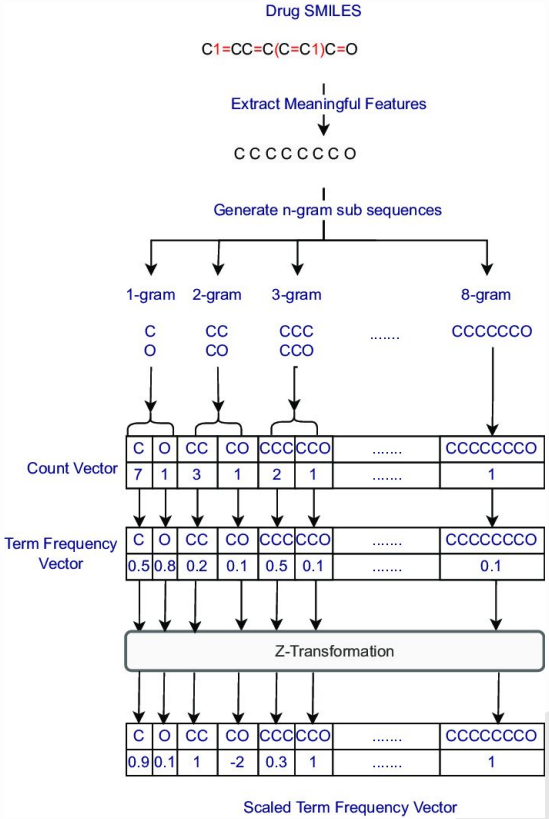
Methods 1.1 Data Collection

- Uniformly sourced from ADCdb (publicly accessible online database)
- All components of the ADC (heavy chain and light chain of the antibody, payload SMILES (Simplified Molecular Input Line Entry System), and linker SMILES) were collected

ADC Name:	Linker Name:	Heavy Chain	Light Chain	Payload Name:
Loncastuximab tesirine	Mal-PEG8-Val-Cit-PABC	Loncastuximab	Loncastuximab	SG3199
Belantamab mafodotin	Maleimido-caproyl	Belantamab	Belantamab	Monomethyl auristatin F
Polatuzumab vedotin	Mc-Val-Cit-PABC	Polatuzumab	Polatuzumab	Monomethyl auristatin E
Disitamab vedotin	Mc-Val-Cit=PABC	Hertuzumab	Hertuzumab	Monomethyl auristatin E
Tisotumab vedotin -tftv	Mc-Val-Cit-PABC	Tisotumab	Tisotumab	Monomethyl auristatin E
Enfortumab vedotin	Mc-Val-Cit-PABC	Enfortumab	Enfortumab	Monomethyl auristatin E
Mirvetuximab soravtansine	Sulfo-SPDB	Mirvetuximab	Mirvetuximab	Mertansine DM4
Gemtuzumab ozogamicin	AcButDMH	Gemtuzumab	Gemtuzumab	N-acetyl-gamma-calichemicin
Sacituzumab govitecan	CL2A	Sacituzumab	Sacituzumab	Active metabolite of irinotecan SN38
Brentuximab vedotin	Mc-Val-Cit-PABC	Brentuximab	Brentuximab	Monomethyl auristatin E
Inotuzumab ozogamicin	AcButDMH	Inotuzumab	Inotuzumab	N-acetyl-gamma-calicheamicin
Moxetumomab pasudotox	Mc-Val-Cit-PABC	Moxetumomab	Moxetumomab	Pseudomonas exotoxin PE38
Cetuximab sarotalocan	Linear alkyl/alkoxy linker	Cetuximab	Cetuximab	IRDye 700DX

Methods 1.2 Feature Representation

- ESM-2 (antibodies)
  - Transformer based language model
  - Output: 1280-dimension feature vectors
- NLP-based method (linkers/payloads)
  - Interprets SMILES as natural language sequences
  - Output: High dimensional feature vector



Methods 1.3 Custom Diffusion Model

- DDPMs (Denoising Diffusion Probabilistic Model)
  - Characterized by its 2 Markov chains for adding/removing noise

- Parameters:
  - **Epochs = 50** *(determined through experimentation and model training results)*
  - Batch size = 32 *(determined through hardware constraints)*
  - **Learning rate = 0.0001** *(determined through experimentation and model training results)*

- Loss function: sums the Mean Squared Error (MSE) & context loss to guide the model's learning process
  - MSE: commonly used for DDPMs because its loss aligns well with properties of Gaussian distributions
  - Context Loss: loss between noise added data and original data

Methods 1.3 Custom Diffusion Model

- Custom Model Key Features:
  - ReLU (rectified linear unit) activation layers: introduces nonlinearity into the model to prevent overfitting
  - Gradient trimming during backpropagation: prevents extreme weight updates, ensuring stable learning

- Linker Generation
  - Input parameters are the other ADC component embeddings for contextual information
  - Denoising process generates the novel linker embeddings

Denoising formula

$$x = \frac{x - \sqrt{1 - \bar{\alpha}_t} \cdot \hat{\epsilon}}{\sqrt{\bar{\alpha}_t}}$$

Methods 1.4 Decoder

- The output of the diffusion model is the linker embeddings which have to be converted back into SMILES strings to be interpreted
- SMILES strings are broken down into basic units to construct a character set dictionary
- Keras-based decoder model
    - High level neural network API commonly used for quickly building and training deep learning models
    - Implements logits (indicates how much model favors specific character as next in a sequence)

Linker Name:	Embeddings	SMILES
Mal-PEG4-N3	[-0.1417271314,0.6906379009,-0.2525271761,-0.0972388806,1.1051467,-0.3876067557,-0.1	[N-]=[N+]=NCOCCOCCOCCOCCC(=O)ON1C(=O)C=CC1=O
Gly5	[-0.1417271314,2.571761511,-0.2525271761,-0.0972388806,1.026263835,-0.3876067557,-0.	C(C(=O)NCC(=O)NCC(=O)NCC(=O)NCC(=O)O)N
Dextran	[-0.1417271314,-2.2641573,-0.2525271761,-0.0972388806,2.547328611,-0.3876067557,-0.1:	C(C1C(C(C(C(O1)OCC2C(C(C(C(O2)OCC(C(C(C(C O)O)O)O)O)O)O)O)
N6-(2-azidoethoxy)carbonyl-L-lysine	[-0.1417271314,2.35372335,-0.2525271761,-0.0972388806,0.1482883664,-0.3876067557,-0.	C(CCNC(=O)OCCN=[N+]=[N-])C[C@@H](C(=O)O)N
Ala-Ala dipeptide	[-0.1417271314,-0.4547221659,-0.2525271761,-0.0972388806,0.7486257697,-0.3876067557	C[C@@H](C(=O)N[C@@H](C)C(=O)O)N
Mc-Val-Ala-PABC	[-0.1417271314,-0.3620070818,-0.2525271761,-0.0972388806,-1.51117671,-0.3876067557,-	C[C@@H](C(=O)NC1=CC=C(C=C1)CO)NC(=O)[C@H](C(C)C)NC(=O)
PEG2-Val-Cit-PABC	[-0.1417271314,-0.1438464722,-0.2525271761,-0.0972388806,-0.9165034358,-0.3876067557	C[C@@H](C(=O)NC1=CC=C(C=C1)CO)NC(=O)[C@H](C(C)C)NC(=O)

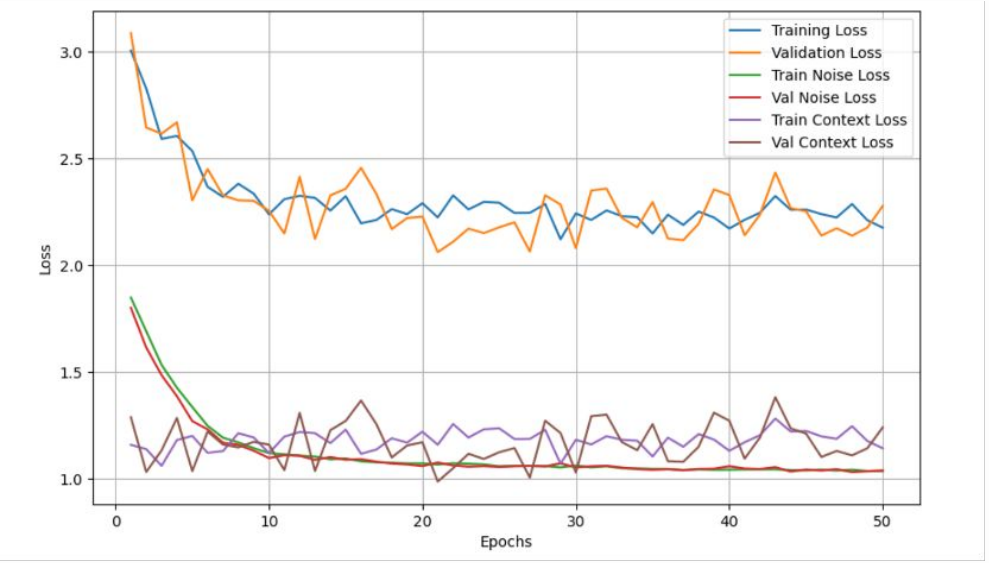
Decoder

- Key features:
    - Dropout layer
    - RDKit to check validity
    - Beam search algorithm

- Parameters:
    - Epochs = 30 *(determined through experimentation and model training results)*
    - Batch size = 32 *(determined through hardware constraints)*

```
SMILES 1: C=CCOCCCCCCCCCCCCCCCCOCCCCCCCCCCCCCCCCCOCCCCCCCCCCCCCCCCCCOCCCCCCCCCCOCCCCCCCCCCCCCCCC is valid.
SMILES 2: C=CCOCCCCCCCCCCCCCCCCOCCCCCCCCCCCCCCCCCOOCCCCCCCCCCCCCCCCCOCCCCCCCCCCOCCCCCCCCCCCCCCCC is valid.
SMILES 3: C=CCOCCCCCCCCCCCCCCCCOCCCCCCCCCCCCCCCCCOCCCCCCCCCCCCCCCCCCCCCCCCCCCCCOCCCCCCCCCCCCCCCC is valid.
SMILES 4: C=CCOCCCCCCCCCCCCCCCCOCCCCCCCCCCCCCCCCCOOCCCCCCCCCCCCCCCCCCCCCCCCCCCCOCCCCCCCCCCCCCCCC is valid.
SMILES 5: C=CCOCCCCCCCCCCCCCCCCOCCCCCCCCCCCCCCCCCOCCCCCCCCCCCCCCCCCCOCCCCCCCCCCOCCCCCCCCCCCCCCCCC is valid.
SMILES 6: CCOCCNCCCCCC(=O)NC(CCCCCCO)CCC(C)C is valid.
SMILES 7: CCCCCCCCCCCOCCCCCCCCCCOCCCCCCCCCCCCNCCCCCCCCCCCCCCCCOCCCCCCC is valid.
SMILES 8: CCCCCCCCCCCOCCCCCCCCCCOCCCCCCCCCCCCNCCCCCCCCCCOCCCCCOCCCCCCC is valid.
SMILES 9: CCCCCCCCCCCOCCCCCCCCCCOCCCCCCCCCOCCNCCCCCCCCCCCCCCCCOCCCCCCC is valid.
SMILES 10: CCCCCCCCCCCOCCCCCCCCCCOCCCCCCCCCCCCNCCCCCCCCCCCCCCOCOCCCCCCC is valid.
SMILES 11: CCCCCCCCCCCOCCCCCCCCCCOCCCCCCCCCCCCNCCCCCCCCCCCCCCCCOCCCCCCC is valid.
SMILES 12: CCCCCCCCCCCOCCCCCCCCCCOCCCCCCCCCCCCNCCCCCCCCCCOCCCCCOCCCCCCC is valid.
SMILES 13: CCCCCCCCCCCOCCCCCCCCCCOCCCCCCCCCOCCNCCCCCCCCCCCCCCCCOCCCCCCC is valid.
SMILES 14: CCCCCCCCCCCOCCCCCCCCCCOCCCCCCCCCCCCNCCCCCCCCCCCCCCOCOCCCCCCC is valid.
SMILES 15: C=C(C)CCCCCCC(=O)CCCCCOCCCCCCCCOCCCCOCOC is valid.
SMILES 16: CCCCCCCCCCCCCCCCCCCCCCCCOOCCCCCCCCCCCCCCCCCCCCC is valid.
SMILES 17: CCCCCCCCCCCCCCCCCCCCCCCOOCCCCCCCCCCCCCCCCCCCCC is valid.
SMILES 18: CCCCCCCCCCCCCCCCCCCCCCOOCCCCCCCOCCCCCCCCCCCCCCC is valid.
SMILES 19: CCCCCCCCCCCCCCCCCCCCCCOOCCCCCCOCCCCCCCCCCCCCCC is valid.
SMILES 20: CCCCCCCCCCCCCCCCCCCCCCOOCCCCCCCCCCCCCCCCCCCCOCC is valid.
SMILES 21: CCCCCCCCC(CC(=O)CCCCCCCCCCCCOCCCCCOCC(O)NO)C(C)C is valid.
SMILES 22: CCCCCCCCC(CC(=O)CCCCCCCCCCCCCCCCCCOCC(O)NO)C(C)C is valid.
SMILES 23: CCCCCCCCC(CC(=O)CCCCCCCCCCCCOCCCCCOCC(C)NO)C(C)C is valid.
SMILES 24: CCCCCCCCC(CC(=O)CCCCCCCCCCCCOCCCCCCCC(O)NO)C(C)C is valid.
SMILES 25: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCOCCCCCCCOCCCOCOOCOO is valid.
SMILES 26: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCOCCCOCCCOCCCOCOOCOO is valid.
SMILES 27: CCCCCCCCCCCCCCCCCCCCCOCCCCCCCCCOCCCCCCCOCCCOCOOCOO is valid.
SMILES 28: CCCCCCCCCCCCCCCCCCCCCOCCCCCCCCCOCCCOCCCOCCCOCOOCOO is valid.
```

Results: Diffusion Model

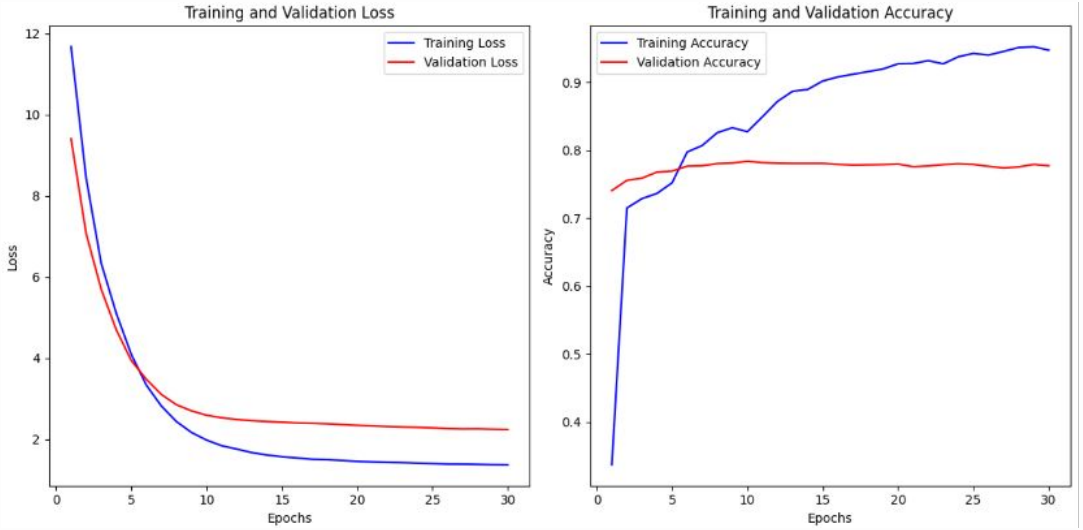


(1) Noise Loss
- Close to a minimum value = strong ability to predict noise
- Minimal gap = good generalization ability
- Smooth curves = stable training process

(2) Context Loss
- Decrease in training and validation = good generalization ability

(3) Problems
- Convergence occurs quickly, suggesting a potential local minimum
- Reasons: small dataset, simple network structure

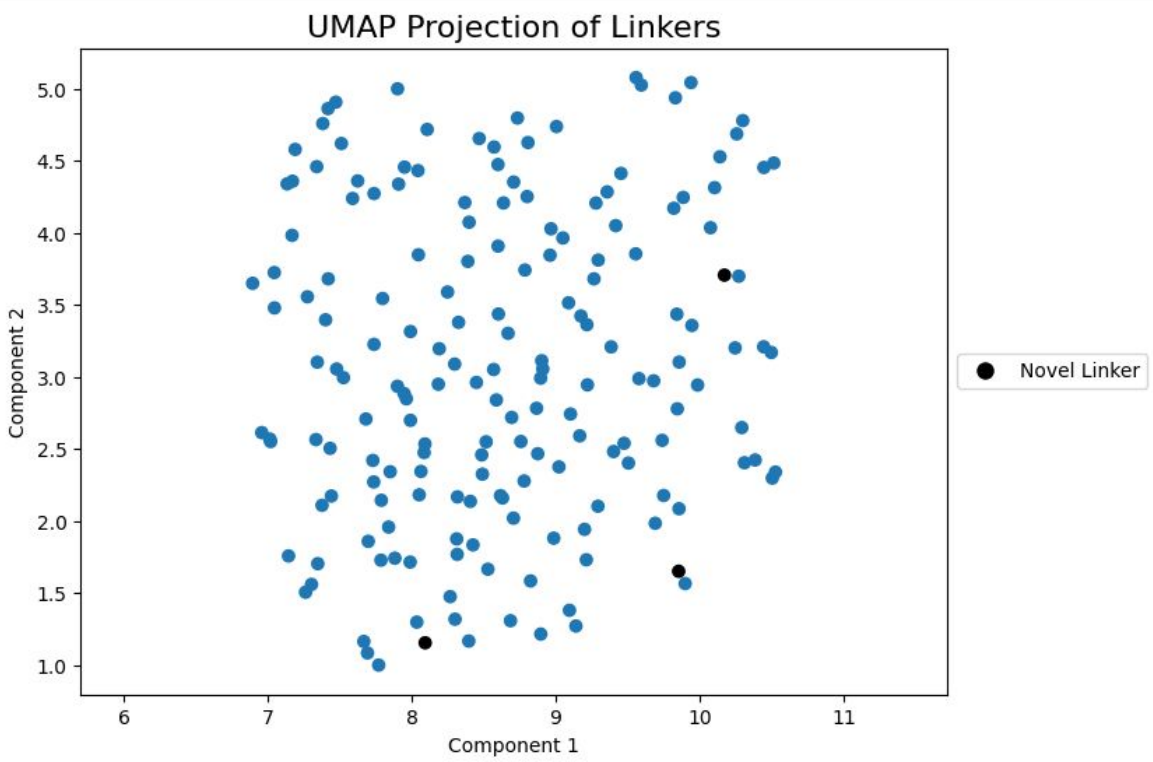
Results: Decoder



(1) Training set loss and validation set loss
- As number of epochs increase, both gradually decrease until a minimum value stabilizes
- Gap between the two increases

(2) Training set accuracy and validation set accuracy
- Both increase with the increasing epochs, final values are close to a maximum
- Final value of validation set accuracy is lower, suggesting some overfitting

UMAP Representation



- Determine nearest neighbours of each novel linker

Linker Evaluation Metrics

ProTox 3.0 and SwissADME

LD50

the dose at which
50% of participants
die upon exposure

Lipophilicity

ability of compound
to dissolve in lipids

Bioavailability

the ability of the drug
to be absorbed/used
by the body

Synthetic Accessibility

how easily it can be built
in a lab

Linker Results

	SMILES	LD50 Value	Toxicity Class	Lipophilicity	Bioavailability Score	Synthetic Accessibility
Thresholds		*LD50>500 mg/kg (moderate toxicity)*	*1-6 (6 is least toxic)*	*LogP 1-3 (favorable lipophilicity)*	*> 0.55 (acceptable bioavailability)*	*SA < 5 (practical synthesis)*
New Linker 3 (molecule 17)	**O=C(NCCN1C(=O)C=CC1=O)CCCSSc1ccccn1**	**790 mg/kg**	**4**	**1.3**	**0.55**	**3.16**
Mal-Me3Lys-Pro	C[N+](C)(C)CCCCC(NC(=O)C1CCCN1)C(=O)NCCNC(=O)CCN1C(=O)C=CC1=O	215 mg/kg	3	-2.29	0.55	4.76
Maleamic methyl ester-based linker 12A	CC(NC(=O)C(NC(=O)CCNC(=O)/C=C/C(=O)O)C(C)C)C(=O)Nc1ccc(CO)cc1	1000 mg/kg	4	0.41	0.11	3.89

Conclusions

- This study demonstrates the potential of diffusion-based generative models to revolutionize the design of linker sequences in ADC
  - Linkers generated are structurally valid, pharmacologically stable, and adapted to requirements of other ADC components
- Significance
  - These advancements hold potential to accelerate preclinical and clinical testing in ADCs by optimizing linker design at the computational stage, reducing costs, and enabling the development of more effective cancer therapies
- Future directions
  - Enriched and improved dataset
  - Increasing complexity of models (ex. Residual diffusion models) to represent the relationship between different ADC components

References

1. World Health Organization. Global cancer burden growing, amidst mounting need for services. (2024)
2. WebMD. Chemotherapy: Types, how it works, procedure and side effects. n.d.
3. Maecker H, Jonnalagadda V, Bhakta S, Jammalamadaka V, & Junutula JR. Exploration of the antibody-drug conjugate clinical landscape. MAbs. 2023; 15(1): 2229101
4. Challener C. Optimization of Linker Chemistries for Antibody-Drug Conjugates. BioPharm International. 2023;36(11):12-15.
5. Ho J, Jain A, & Abbeel P. Denoising Diffusion Probabilistic Models. arXiv. 2020
6. Baah S, Laws M, & Rahman K. M. Antibody-Drug Conjugates-A Tutorial Review. Molecules (Basel, Switzerland). 2021;26(10):2943.
7. Shen L. et al. ADCdb: the database of antibody-drug conjugates. Nucleic Acids Res., (2023).
8. Lin Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379:1123-1130.
9. Sharma R, Saghapour E, & Chen J. Y. An NLP-based technique to extract meaningful features from drug SMILES. iScience. 2024;27(3)
10. Banerjee P, Kemmler, E, Dunkel, M, & Preissner R. ProTox 3.0: a webserver for the prediction of toxicity of chemicals. Nucleic Acids Res., 2024.
11. Daina A, Michielin O, Zoete V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci. Rep. 2017;7:42717

Thanks for listening!

CREDITS: This presentation template was created by **Slidesgo,** and includes icons, infographics & images by **Freepik**