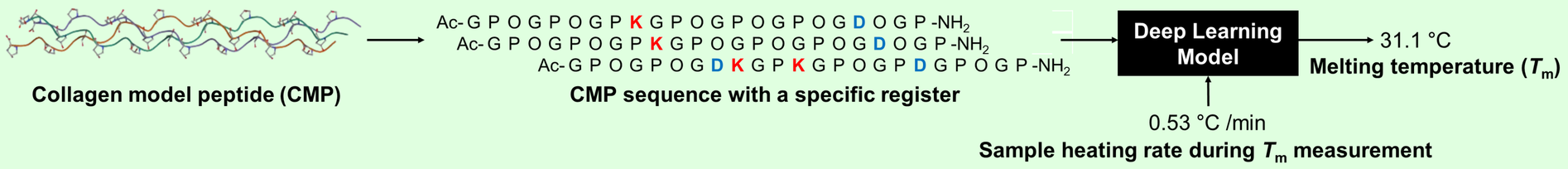
# A Deep Learning Model for Predicting the Thermal Stability of Collagen Triple Helices

Kiseop Im, Ian Warm, Tomas Fiala,* and Helma Wennemers*

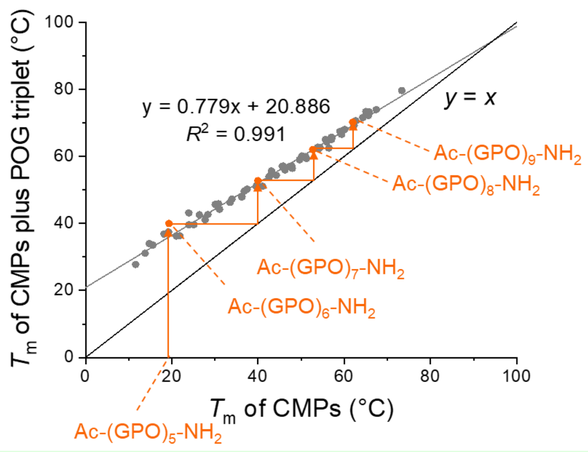
Laboratory of Organic Chemistry, ETH Zürich, D-CHAB, Vladimir-Prelog-Weg 3, 8093 Zürich, Switzerland.

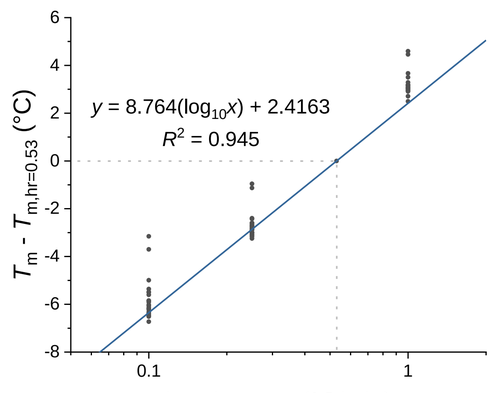
## 1. Introduction

Collagen model peptides (CMPs) are key tools for investigating the physico-chemical properties of collagen, the most abundant mammalian protein, as well as designing synthetic collagen-based materials and probes.[1] Like collagen, CMPs self-assemble into a triple-helical structure that is characterized by its melting temperature ($T_m$).[2,3] The $T_m$ is highly dependent on the CMP sequence.[1,4] Because of the colossal number of possible sequence variations, computational methods are necessary for deciphering the trends in the thermal stability of collagen triple helices. Despite several attempts to predict $T_m$ values from CMP sequences,[5-7] no computational models integrate all key aspects that influence the $T_m$ to achieve accurate predictions. Here, we design a deep-learning model that dissects the pairwise intra- and interstrand interactions in CMPs and uses them to accurately predict $T_m$ values from CMP sequences. Our approach comprehensively deals with critical aspects affecting the experimental $T_m$ such as frame shifts,[8] capping groups,[9] and heating rates.




## 2. Considered Trends in $T_m$



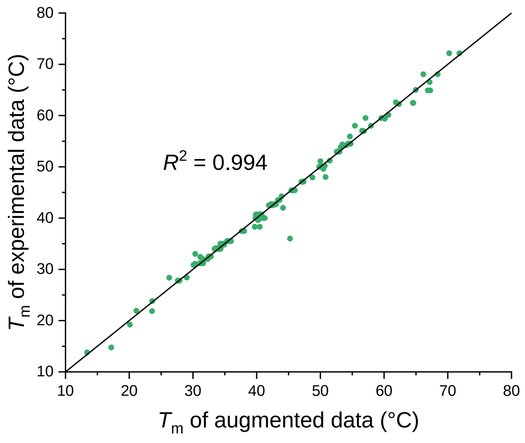




- Any sequence modifications or pairwise-interaction additions change the $T_m$ of the CMP.
- The change in $T_m$ linearly depends on the basal $T_m$ of the parental sequence but is not affected by the existing sequence or pairwise interactions.
- $T_m$ increases with a higher heating rate logarithmically

| Sequence | $T_m$ (°C) |
|---|---|
| H-GPOGPOGPOGPOGPOGPOGPOGPO-OH | 27.8 |
| Ac-GPOGPOGPOGPOGPOGPOGPOGPO-OH | 35.6 |
| Ac-GPOGPOGPOGPOGPOGPOGPOGPO-$NH_2$ | 52.7 |
| Ac-GPOGPOGPOGPOGPOGPOGPOGPOG-$NH_2$ | 45.5 |
| Ac-GPOGPOGPOGDRGPOGPOGPOGPOGG-$NH_2$ | 37.1 |

Existing sequence-$T_m$ pairs

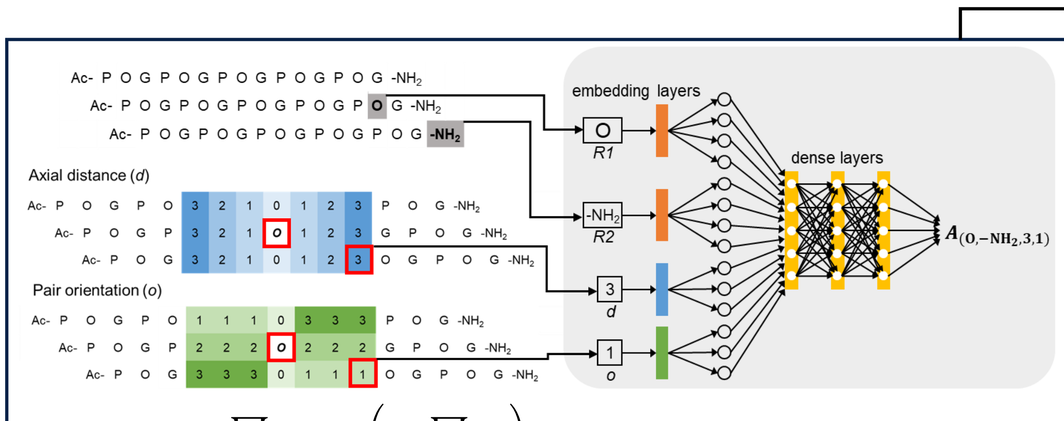
Augmented pairs for decoupling




- Using previously identified general trends,[8,9] 3,695 more sequence-$T_m$ pairs were mathematically generated.
- 214 augmented data cross-validated with existing experimental data

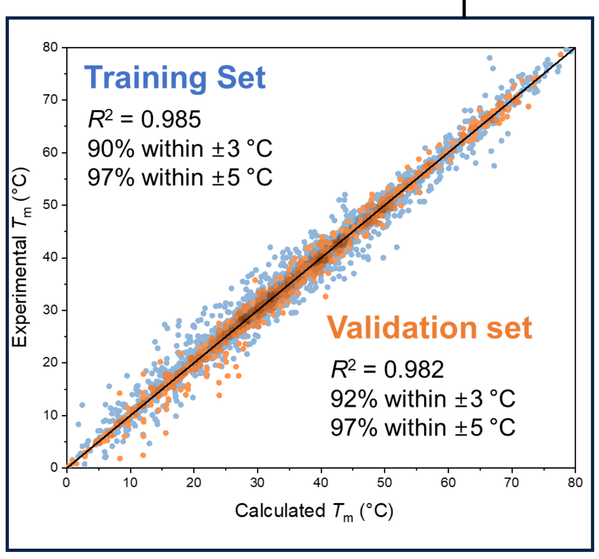
## 3. Model Training and Testing

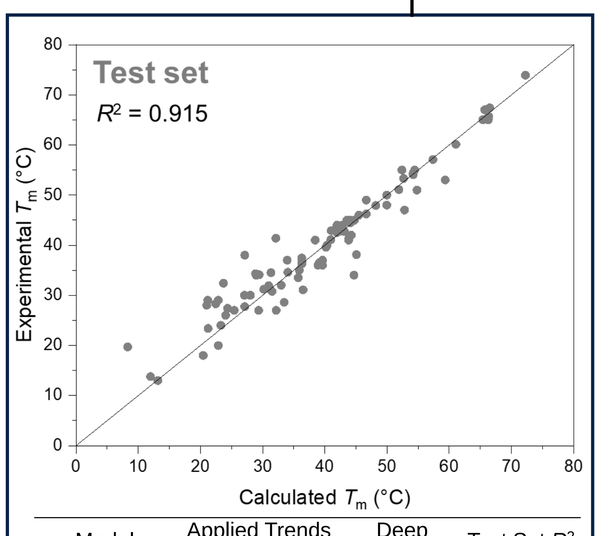
1,099 CMP-$T_m$ pairs from 84 references → 4,694 CMP-$T_m$ pairs through data augmentation → dataset split according to the uniqueness of the data points → deep-learning model construction → training and optimization → model testing




$$\therefore T_{m,pred} = T_0 \prod_{i \in PI} A_i^{n_i} + T_f \left(1 - \prod_{i \in PI} A_i^{n_i}\right) + \alpha(\ln(\text{heating rate}) - \ln(0.53\,^\circ\text{C/min}))$$

- The model calculates the contribution of pairwise interactions in the CMP to the $T_m$.




| Model | Applied Trends in $T_m$ | Deep Learning | Test Set $R^2$ |
|---|---|---|---|
| Model A[5] | O | X | N/A |
| **This work** | **O** | **O** | **0.92** |
| Model B[6] | X | O | N/A |
| Model C[7] | X | O | 0.28 |
| Model D[7] | X | O | 0.62 |

## 4. Conclusion

- We constructed a deep learning model that considers key trends in collagen $T_m$ and successfully trained the model with reported and augmented data.
- The deep learning model achieves unprecedented accuracy and generalization in predicting collagen $T_m$ values.
- **Significance:** our model represents a general tool that will contribute to a better understanding of CMP sequence-property relationships and aid the design of collagen triple helices with desired stability.

Partner/Sponsor: FNSNF Swiss National Science Foundation; The Branco Weiss Fellowship Society in Science

## References

1. Shoulders, M. D.; Raines, R. T., Collagen Structure and Stability. *Annu. Rev. Biochem.* **2009**, *78* (1), 929-958.
2. Okuyama, K., Miyama, K., Mizuno, K., & Bächinger, H. P. Crystal structure of (Gly-Pro-Hyp) 9: Implications for the collagen molecular model. *Biopolymers*, **2012**, *97*(8), 607-616.
3. Privalov, P. Stability of proteins: proteins which do not present a single cooperative system. *Advances in protein chemistry.* **1982**, *35*, 1-104.
4. Persikov, A. V., Ramshaw, J. A., Kirkpatrick, A. & Brodsky, B. Amino acid propensities for the collagen triple-helix. *Biochemistry.* **2000**, *39*, 14960-14967.
5. Walker, D. R. *et al.* Predicting the stability of homotrimeric and heterotrimeric collagen helices. *Nature Chemistry.* **2021,** *13,* 260-269.
6. Yu, C.-H. *et al.* ColGen: An end-to-end deep learning model to predict thermal stability of de novo collagen sequences. *Journal of the mechanical behavior of biomedical materials.* **2022,** *125*, 104921.
7. Khare, E., Gonzalez-Obeso, C., Kaplan, D. L. & Buehler, M. J. CollagenTransformer: end-to-end transformer model to predict thermal stability of collagen triple helices using an NLP approach. *ACS Biomaterials Science & Engineering.* **2022**, *8,* 4301-4310.
8. Fiala, T.; Barros, E. P.; Ebert, M.-O.; Ruijsenaars, E.; Riniker, S.; Wennemers, H., Frame Shifts Affect the Stability of Collagen Triple Helices. *J. Am. Chem. Soc.* **2022**, *144* (40), 18642-18649.
9. Fiala, T.; Barros, E. P.; Heeb, R.; Riniker, S.; Wennemers, H., Predicting Collagen Triple Helix Stability through Additive Effects of Terminal Residues and Caps. *Angew. Chem., Int. Ed.* **2023**, 62 (3), e202214728