Geometric Epitope and Paratope Prediction

Marco Pegoraro[*] PEGORARO@DI.UNIROMA1.IT
Sapienza, University of Rome

Clémentine Dominé[*] CLEMENTINE.DOMINE.20@UCL.AC.UK
Gatsby Computational Neuroscience Unit, University College London

Emanuele Rodolà RODOLA@DI.UNIROMA1.IT
Sapienza, University of Rome

Petar Veličković PETARV@GOOGLE.COM
Google DeepMind

Andreea Deac DEACANDR@MILA.QUEBEC
SMila, Université de Montréal

Abstract

Antibody-antigen interactions play a crucial role in identifying and neutralizing harmful foreign molecules. In this paper, we investigate the optimal representation for predicting the binding sites in the two molecules and emphasize the importance of geometric information. Specifically, we compare different geometric deep learning methods applied to proteins' inner (I-GEP) and outer (O-GEP) structures. We incorporate 3D coordinates and spectral geometric descriptors as input features to fully leverage the geometric information. Our research suggests that surface-based models are more efficient than other methods, and our O-GEP experiments have achieved state-of-the-art results with significant performance improvements.

1. Introduction

Identifying the binding sites of antibodies is essential for developing vaccines and synthetic antibodies. These binding sites, called paratopes, can bind to antigens, wherein the corresponding binding site is known as the epitope, thus neutralizing harmful foreign molecules in the body. Experimental methods for determining the residues that belong to the paratope and epitope are time-consuming and expensive, highlighting the need for computational tools to facilitate the rapid development of therapeutics.

The shape and structure of molecules play a crucial role in determining their interactions with other molecules, as complementary geometric shapes are required for successful binding (Fischer, 1894). Various approaches have been taken in the literature to address the task of epitope and paratope prediction, including sequential (Liberis et al., 2018; Deac et al., 2019) and structural (Krawczyk et al., 2014; Del Vecchio et al., 2021) methods. Furthermore, Geometric deep learning has emerged as a powerful tool for predicting protein-protein interactions (Isert et al., 2023), with graph-based representations being one of the most common approaches Tubiana et al. (2022); Stärk et al. (2022). These methods leverage the geometric information of the molecules to learn complex relationships between epitopes and paratopes. For instance, (Del Vecchio et al., 2021) and da Silva et al. (2022) use the graph

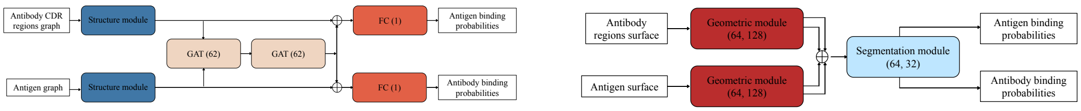
[*] Equal Contribution

structure to compute features based on neighbouring residues, which are then aggregated to highlight the most probable region of interaction. An alternative approach is to represent proteins as surfaces. MaSIF (Gainza et al., 2020) focuses on the more general problem of protein interaction region prediction and uses a surface representation learned through convolutions defined on the surface. PiNet (Dai and Bailey-Kellogg, 2021) represents the protein surface as a point cloud and employs PointNet (Qi et al., 2017) to classify points as interacting or not. On the contrary, Zhang et al. (2023) model the surface of a molecule as a graph and apply an equivariant graph neural network (EGNN, (Satorras et al., 2021)) for binding site prediction. Integrating structural and geometric information has proven to be a promising approach for improving protein interaction prediction. Still, few studies have focused on the specific case of epitope and paratope prediction (Cia et al., 2023).

Our approach, GEP (Geometric Epitope-Paratope) Prediction, proposes different geometric representations of the molecules to create accurate predictors for predicting antibody-antigen binding sites. Our paper introduces several contributions, including the analysis of the importance of geometric information within graph learning using equivariant layers for improved predictions. Moreover, we fully leverage molecular geometric information by representing molecules as surfaces and employing spectral geometry techniques, leading to state-of-the-art performance. Additionally, we will provide a dataset generation pipeline for PDB molecules, offering molecular representations in both graph and surface formats, facilitating comprehensive cross-method comparisons. The code is publically available.

2. Method

In our experiments, we considered two scenarios: a protein represented through its inner structure (*I-GEP*) and outer structure (*O-GEP*). Details on the data and how we construct the different representations for each model are reported in Appendix A.


Antibody CDR regions graph
Structure module
GAT (62)
GAT (62)
FC (1)
Antigen binding probabilities
Antigen graph
Structure module
FC (1)
Antibody binding probabilities
Antibody regions surface
Geometric module (64, 128)
Segmentation module (64, 32)
Antigen binding probabilities
Antigen surface
Geometric module (64, 128)
Antibody binding probabilities

(a) I-GEP

(b) O-GEP

Figure 1: Our model architecture is represented with arrows indicating data flow between modules, using color-coded blocks to represent layers or modules, with text inside each block specifying the layer type. The model takes antibody-antigen pairs as input, featuring node-level features for IGEP and surface point-level features for OGEP, and produces binding probabilities for each input node or point.

I-GEP Our I-GEP model is a method for predicting epitopes and paratopes using a graph-based approach that captures the inner structure of a protein. The I-GEP model has two main components: a structural module that computes an embedding for each residue using the graph structure and a graph attention network (GAT) that combines information from both the antigen and antibody residues. The network then predicts both epitope and paratope residues simultaneously using a fully connected layer, as shown in Fig. 1(a). To improve the accuracy of our predictions, we integrate geometric information into the I-GEP model using two different approaches. In the first approach, EPMP$_{xyz}$, we use graph convolutional network layers in the structural module as in EPMP Del Vecchio et al. (2021), but we include the centred 3D coordinates of residues in the input features. The second approach, $E(n)$-*EPMP*, uses the $E(n)$ invariant layer encoder from EGNN Satorras et al. (2021) instead of graph convolutional networks. This approach considers only the distances between residues, making it invariant to translations, rotations, and reflections on the residue positions in each molecule.

O-GEP Our O-GEP model operates on the protein's surface and includes a geometric module that uses the surface's geometry to spread information across it. This process generates features that are then combined and shared between the antibody and antigen through fully connected layers (segmentation module), resulting in an interaction probability for each point on the surface, as shown in Fig. 1(b).

We explore two different models for the geometric module. As a baseline, we use PointNet (Qi et al., 2017) to recreate the architecture proposed in PiNet Dai and Bailey-Kellogg (2021). The second model employs diffusion layers from DiffNet (Sharp et al., 2022) to propagate features on the surface. This makes our model robust against surface perturbations and suitable for handling meshes and point clouds with fewer points.

We further examine the impact of using the Heat Kernel Signature (HKS) as an extra geometric descriptor input. The HKS (Sun et al., 2009) is a concise point-wise spectral signature which summarizes local and global information about the intrinsic geometry of a shape by capturing the properties of the heat diffusion process on the surface. One of the key benefits of using HKS is that it remains stable even under minor surface perturbations, thus enabling it to withstand even conformational rearrangements of the proteins. To utilize the HKS descriptor, we concatenate it with the input features at each point on the surface and then pass the concatenated data through the geometric module.

To transfer the binding probabilities from the protein's surface to the residues, we utilized the average of all the points on the surface that correspond to the same residues. This method ensures that the binding probabilities are accurately represented in the residue space, enabling us to make reliable predictions about epitope and paratope locations.

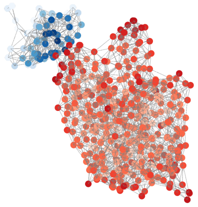
Training and evaluation The networks were trained using the class-weighted binary cross-entropy loss and the Adam SGD optimizer to handle imbalanced binary classification tasks. We report training details in Appendix B. Given the significant disparity in class sizes, we utilize Matthew's correlation coefficient (MCC) between the residues' classification as our main benchmarking metric for model evaluation. We also report the area under the receiver operating characteristic curve (AUC ROC) and the area under the precision-recall curve (AUC PR) as used in Dai and Bailey-Kellogg (2021) and Del Vecchio et al. (2021). All reported values are aggregated across five random seeds to ensure the robustness of our findings.

3. Results

In this section, we report the results of our experiments and demonstrate the contribution of geometric information on the task of epitope-paratope prediction.

	Antigen			Antibody		
	MCC	AUC ROC	AUC PR	MCC	AUC ROC	AUC PR
EPMP	0.09 ± 0.01	0.61 ± 0.01	0.12 ± 0.00	0.39 ± 0.02	0.79 ± 0.01	0.53 ± 0.01
EPMP$_{xyz}$	0.10 ± 0.01	0.63 ± 0.01	0.15 ± 0.01	0.38 ± 0.02	0.79 ± 0.01	0.53 ± 0.01
$E(n)$-EPMP	$\mathbf{0.14 \pm 0.01}$	$\mathbf{0.68 \pm 0.02}$	$\mathbf{0.16 \pm 0.01}$	$\mathbf{0.44 \pm 0.11}$	$\mathbf{0.82 \pm 0.07}$	$\mathbf{0.60 \pm 0.10}$

(b) Quantitative results as mean and standard deviation (±)

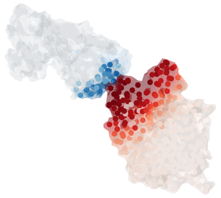


(c) Qualitative example

(a) Results from I-GEP models.

	Antigen			Antibody		
	MCC	AUC ROC	AUC PR	MCC	AUC ROC	AUC PR
PiNet (xyz)	0.39 ± 0.05	0.89 ± 0.01	0.44 ± 0.02	0.26 ± 0.12	0.77 ± 0.03	0.52 ± 0.08
PiNet (xyz+hks)	0.30 ± 0.04	0.87 ± 0.02	0.37 ± 0.06	0.22 ± 0.05	0.74 ± 0.00	0.47 ± 0.02
DiffNet$_{pc}$ (xyz)	0.41 ± 0.06	$\mathbf{0.90 \pm 0.01}$	0.49 ± 0.02	0.30 ± 0.06	0.79 ± 0.01	0.56 ± 0.03
DiffNet$_{pc}$ (hks)	0.07 ± 0.05	0.66 ± 0.02	0.14 ± 0.01	0.44 ± 0.03	$\mathbf{0.85 \pm 0.00}$	0.68 ± 0.01
DiffNet$_{pc}$ (xyz+hks)	$\mathbf{0.44 \pm 0.03}$	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.50 \pm 0.02}$	0.23 ± 0.06	0.77 ± 0.04	0.51 ± 0.05
DiffNet$_{m}$ (xyz)	0.42 ± 0.03	$\mathbf{0.90 \pm 0.01}$	0.48 ± 0.05	0.24 ± 0.08	0.78 ± 0.02	0.52 ± 0.03
DiffNet$_{m}$ (hks)	0.09 ± 0.02	0.64 ± 0.02	0.14 ± 0.01	$\mathbf{0.49 \pm 0.01}$	$\mathbf{0.85 \pm 0.00}$	$\mathbf{0.69 \pm 0.01}$
DiffNet$_{m}$ (xyz+hks)	0.42 ± 0.06	$\mathbf{0.90 \pm 0.01}$	0.46 ± 0.07	0.28 ± 0.06	0.77 ± 0.02	0.52 ± 0.04

(e) Quantitative results as mean and standard deviation (±)



(f) Qualitative example

(d) Results from O-GEP models.

Figure 2: **Left**: Quantitative results evaluated on the residues. We report the Matthew's correlation coefficient (MCC), area under the receiver operating characteristic curve (AUR ROC), AUC PR the area under the precision-recall curve (AUC PR). We write in bold the best results. **Right**: Representation of binding prediciton on the antibody-antigen complex number '4jr9'. The continuous binding predictions are represented as a color gradient in blue and red for the antigen and antibody, respectively.

I-GEP results We conducted experiments to evaluate the effectiveness of incorporating geometric information by comparing our proposed models from Section 2 with the EPMP model proposed in Del Vecchio et al. (2021). Our results, presented in Table 3(b), demonstrate that the inclusion of geometric information leads to a meaningful increase in performance. Specifically, the use of the $E(n)$ invariant layer ($E(n)$-EPMP) resulted in an improvement in all metrics for both antibody and antigen.

O-GEP results To test the performance of O-GEP models, we consider the methods proposed in Section 2 with different combinations of input features. In addition to the physicochemical features, we test different combinations of geometric information: 3d coordinates (xyz) and Heat Kernel Signature (HKS). For the DiffNet models, we consider both the point cloud ($_{pc}$) and the mesh ($_{m}$) of the surface. The results are summarized in Table 3(e). Incorporating diffusion layers (DiffNet) along with 3D coordinates and Heat Kernel Signature as additional features consistently outperformed the baseline method PiNet. The use of these techniques led to an MCC score twice as high as that obtained by the I-GEP models. However, unlike epitope prediction, the paratope prediction did not show the same level of improvement with O-GEP models. In this case, the best results were achieved by considering only the HKS features and diffusion layers. In Appendix C, we also show the metrics computed only on residues with a representing point on the surface.

Qualitative results We plot the binding probability on the residuals computed by the models as increasing intensity colours. Figure 3(c) shows the results of the $E(n)$-*EPMP* on the residual graph. The epitope prediction focuses on sparse regions of the antigene, such as the spiky edges. In contrast, paratope prediction concentrates on the residues closest to the antigen. In Figure 3(f), the predictions of $\textsc{DiffNet}_{pc}$ $_{\text{(XYZ+HKS)}}$ are shown on both the surface and residues of the molecules. The predictions are highly localized on the region nearest to the binding molecule. It's worth noticing that the 3d coordinates given as input to the models are centred and randomly rotated, providing no prior knowledge of the binding region.

4. Conclusions

We investigated the effectiveness of geometric deep learning techniques in predicting antibody-antigen interactions. Our results indicate that incorporating geometric information is crucial for accurately predicting epitope and paratope regions. Specifically, the use of invariant representation in I-GEP models outperformed previous models, and O-GEP models with diffusion layers and additional geometric features achieved state-of-the-art performance. Our study highlights the potential of geometric deep learning in computational biology. Future research could explore using spectral shape analysis to address the more complex problem of conformational rearrangement in antigen-antibody binding (Stanfield et al., 1994).

Acknowledgments

This work is supported by the ERC grant no.802554 (SPECGEO), PRIN 2020 project no.2020TA3K9N (LEGO.AI), and PNRR MUR project PE0000013-FAIR. C.D. was supported by the Gatsby Charitable Foundation (GAT3755) and G-research grants for PhD students.

References

Gabriel Cia, Fabrizio Pucci, and Marianne Rooman. Critical review of conformational b-cell epitope prediction methods. *Briefings in Bioinformatics*, 24(1):bbac567, 2023.

Bruna Moreira da Silva, YooChan Myung, David B Ascher, and Douglas EV Pires. epitope3d: a machine learning method for conformational b-cell epitope prediction. *Briefings in Bioinformatics*, 23(1):bbab423, 2022.

Bowen Dai and Chris Bailey-Kellogg. Protein interaction interface region prediction by geometric deep learning. *Bioinformatics*, 37(17):2580–2588, 2021.

Andreea Deac, Petar Veličković, and Pietro Sormanni. Attentive cross-modal paratope prediction. *Journal of Computational Biology*, 26(6):536–545, 2019.

Alice Del Vecchio, Andreea Deac, Pietro Liò, and Petar Veličković. Neural message passing for joint paratope-epitope prediction. *arXiv preprint arXiv:2106.00757*, 2021.

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M Deane. Sabdab: the structural antibody database. *Nucleic acids research*, 42(D1):D1140–D1146, 2014.

Emil Fischer. Einfluss der configuration auf die wirkung der enzyme. *Berichte der deutschen chemischen Gesellschaft*, 27(3):2985–2993, 1894.

Pablo Gainza, Freyr Sverrisson, Frederico Monti, Emanuele Rodola, D Boscaini, MM Bronstein, and BE Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, 2020.

Clemens Isert, Kenneth Atz, and Gisbert Schneider. Structure-based drug design with geometric deep learning. *Current Opinion in Structural Biology*, 79:102548, 2023.

Konrad Krawczyk, Xiaofeng Liu, Terry Baker, Jiye Shi, and Charlotte M Deane. Improving b-cell epitope prediction and its application to global antibody-antigen docking. *Bioinformatics*, 30(16):2288–2294, 2014.

Edgar Liberis, Petar Veličković, Pietro Sormanni, Michele Vendruscolo, and Pietro Liò. Parapred: antibody paratope prediction using convolutional and recurrent neural networks. *Bioinformatics*, 34(17):2944–2950, 2018.

Jens Meiler, Michael Müller, Anita Zeidler, and Felix Schmäschke. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Molecular modeling annual*, 7(9):360–369, 2001.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pages 9323–9332. PMLR, 2021.

Nicholas Sharp, Souhaib Attaiki, Keenan Crane, and Maks Ovsjanikov. Diffusionnet: Discretization agnostic learning on surfaces. *ACM Transactions on Graphics (TOG)*, 41(3): 1–16, 2022.

Robyn L Stanfield, Midori Takimoto-Kamimura, James M Rini, Albert T Profy, and Ian A Wilson. Major antigen-induced domain rearrangements in an antibody. *Structure*, 1(2): 83–93, 1994.

Hannes Stärk, Octavian Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. Equibind: Geometric deep learning for drug binding structure prediction. In *International Conference on Machine Learning*, pages 20503–20521. PMLR, 2022.

Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum*, volume 28, pages 1383–1392. Wiley Online Library, 2009.

Jérôme Tubiana, Dina Schneidman-Duhovny, and Haim J Wolfson. Scannet: an interpretable geometric deep learning model for structure-based protein binding site prediction. *Nature Methods*, 19(6):730–739, 2022.

Thom Vreven, Iain H Moal, Anna Vangone, Brian G Pierce, Panagiotis L Kastritis, Mieczyslaw Torchala, Raphael Chaleil, Brian Jiménez-García, Paul A Bates, Juan Fernandez-Recio, et al. Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *Journal of molecular biology*, 427 (19):3031–3041, 2015.

Yang Zhang, Wenbing Huang, Zhewei Wei, Ye Yuan, and Zhaohan Ding. Equipocket: an e (3)-equivariant geometric graph neural network for ligand binding site prediction. *arXiv preprint arXiv:2302.12177*, 2023.

Appendix A. Data representation

Comparing methods across different molecular representations is crucial for advancing research in molecular modelling. We developed a reusable pipeline that generates a dataset to evaluate methods using inner and outer structure representations. We collected a dataset of 133 protein complexes from Epipred Krawczyk et al. (2014), with 103 for training and 30 for testing. The training and test sets have been selected to share no more than 90% pairwise sequence identity. The PDB files were obtained from the Sabdab database Dunbar et al. (2014). In the test set, 7.8% of antigen residues were labelled as positive. Additionally, we used a separate set of 27 protein complexes from PECAN derived from a subset of the Docking Benchmark v5 Vreven et al. (2015) to validate our results.

For each protein, we construct a residue graph (Figure 3(c)), representing residues as nodes and establishing edges between the 15 nearest neighboring residues within a 10 Å radius. Each residue is characterized by a 28-dimensional physicochemical feature vector. This vector encompasses a one-hot encoding of the amino acid, encompassing 20 possible types along with one for an unclassified type. Additionally, seven other features are included that portray the physical, chemical, and structural attributes of the amino acid type. These supplementary features can be viewed as a consistent embedding, as outlined in Meiler et al. (2001).

For each protein, we generated a surface mesh (Figure 3(f)) using the PyMOL API with a 1.4 Å water probe radius. We associated each point on the protein's surface with a residue by finding the closest atom to that point. This association was then used to transfer the feature of each residue to the points on the surface.

Table 1: I-GEP quantitative results evaluated on the surface residues. We report the mean and standard deviation (±) over multiple runs.

	Antigen			Antibody		
	MCC	AUC ROC	AUC PR	MCC	AUC ROC	AUC PR
EPMP	0.08 ± 0.01	0.58 ± 0.01	0.13 ± 0.00	0.33 ± 0.03	0.74 ± 0.01	0.56 ± 0.01
EPMP$_{xyz}$	0.08 ± 0.01	0.60 ± 0.01	**0.16** ± 0.01	0.33 ± 0.02	0.74 ± 0.01	0.56 ± 0.01
$E(n)$-EPMP	**0.11** ± 0.01	**0.64** ± 0.01	**0.16** ± 0.01	**0.39** ± 0.11	**0.78** ± 0.07	**0.63** ± 0.08

Appendix B. Hyper-parameters

During training, we combined the losses from both tasks, paratope and epitope prediction. To enhance model robustness, we applied random rotations to dataset instances. Hyperparameter tuning involved a search for the optimal learning rate from the set $\{10^{-2}, 10^{-3}, 5 \times 10^{-3}, 10^{-5}\}$ and kept the model with the best performance on the validation set. . After the hyperparameter search, we found that the best learning rates were: 10^{-3} for EPMP and PiNet, 10^{-2} for $E(n)$-EPMP, 5×10^{-3} for DiffNet. All models were trained for 200 epochs to ensure validation loss saturation, and the weights yielding the best validation metrics during training were selected. We conducted training with five random seeds for each model, evaluating performance using the weights yielding the best validation set results in each run.

The surface generated by PyMOL is composed of around 14k points. To ease and fast the training procedure we subsampled the surface considering only 2k points. In the case of point clouds, we used a random subsampling during training, while for the mesh we used a simplification method base on quadric error metrics.

B.1. Layer dimensions

For the EPMP$_{xyz}$ model, we use a graph convolution layer with inner dimension 31 and two GAT layers with inner dimension 62. In contrast, for the $E(n)$-EPMP, we use one $E(n)$-invariant layer with an inner dimension of 28 and two GAT layers with inner dimension 56.

For all the O-GEP models, the geometric module comprises two layers with dimensions 64 and 128, while the segmentation module is composed of two layers with dimensions 64 and 32.

Appendix C. Outer residues

The outer representation can't include the inner residues because they are too far from the protein's surface representation. As a result, the O-GEP model can't predict those residues. To see how this affects the predictions, we show the results for both I-GEP and O-GEP in Table 1, considering only the outer residues represented by the surface.

Table 2: O-GEP quantitative results evaluated on the surface residues. We report the mean and standard deviation (±) over multiple runs.

	Antigen			Antibody		
	MCC	AUC ROC	AUC PR	MCC	AUC ROC	AUC PR
PiNet (xyz)	0.38 ± 0.04	0.87 ± 0.01	0.45 ± 0.02	0.26 ± 0.12	0.77 ± 0.03	0.52 ± 0.08
PiNet (xyz+hks)	0.29 ± 0.05	0.84 ± 0.02	0.37 ± 0.04	0.13 ± 0.06	0.64 ± 0.01	0.47 ± 0.04
DiffNet$_{pc}$ (xyz)	0.40 ± 0.05	**0.88** ± 0.01	0.49 ± 0.02	0.26 ± 0.06	0.71 ± 0.02	0.56 ± 0.04
DiffNet$_{pc}$ (hks)	0.05 ± 0.04	0.58 ± 0.03	0.14 ± 0.01	0.40 ± 0.02	**0.81** ± 0.01	0.69 ± 0.01
DiffNet$_{pc}$ (xyz+hks)	**0.43** ± 0.03	**0.88** ± 0.01	**0.50** ± 0.02	0.19 ± 0.05	0.68 ± 0.06	0.51 ± 0.05
DiffNet$_{m}$ (xyz)	0.41 ± 0.03	**0.88** ± 0.01	0.49 ± 0.05	0.20 ± 0.07	0.69 ± 0.03	0.53 ± 0.03
DiffNet$_{m}$ (hks)	0.05 ± 0.01	0.56 ± 0.02	0.14 ± 0.01	**0.43** ± 0.02	0.80 ± 0.01	**0.70** ± 0.01
DiffNet$_{m}$ (xyz+hks)	0.41 ± 0.06	**0.88** ± 0.02	0.46 ± 0.07	0.23 ± 0.06	0.68 ± 0.04	0.52 ± 0.04