IOCC 2026 Conference
The 5th International Online Conference on Crystals
15–17 June 2026 | Online
MDPI

Identifying Key Drivers of Nanobody Crystallization using Machine Learning

Kaisar Ahmad Sheikh[1], Fermin Otalora Muñoz[1], José A. Gavira[1*]
[1]Laboratorio de Estudios Cristalográficos-IACT-CSIC, Armilla, Granada, Spain, 18100
kaisarahmadsheikh@gmail.com; f.otalora@csic.es ; j.gavira@csic.es

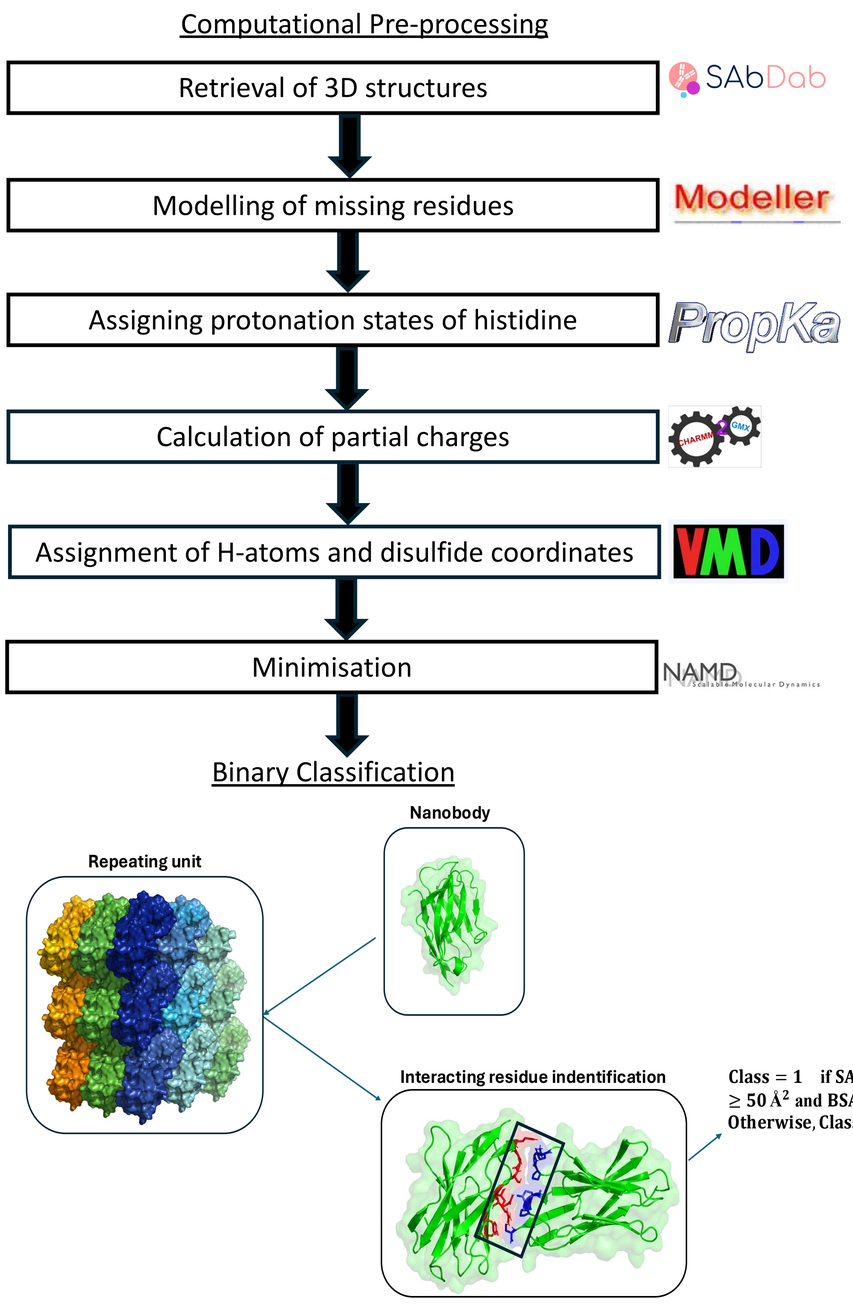
INTRODUCTION

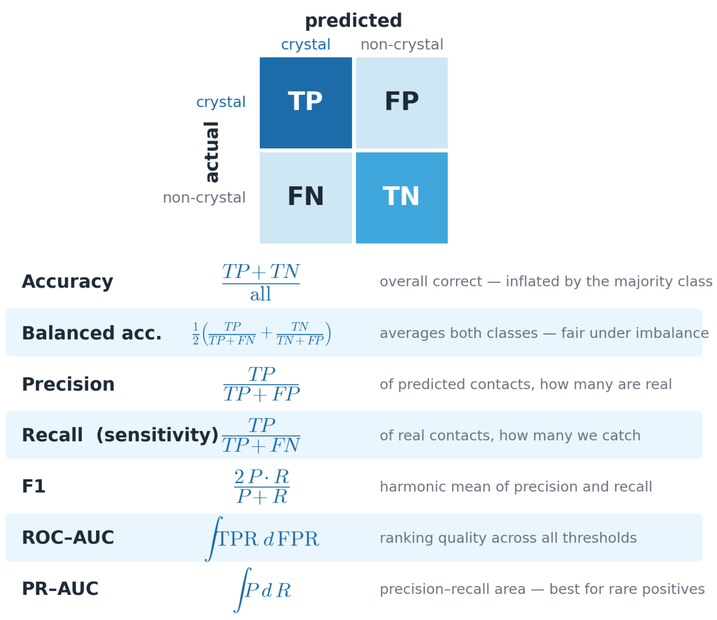
Antibody-based therapeutics are essential biopharmaceuticals, yet their crystallization—crucial for structural characterization, downstream processing, and advanced formulation—remains severely hindered by their large size, conformational flexibility, and a scarcity of structural data. To overcome this bottleneck, this study aims to decode the residue-level determinants of crystal formation and establish a machine learning framework to rationally predict crystallization-enhancing mutations. Using nanobodies as a highly tractable proof-of-concept system, we curated monomeric structures to classify interface residues as crystal-site or non-crystal-site based on different residue level decsriptors.

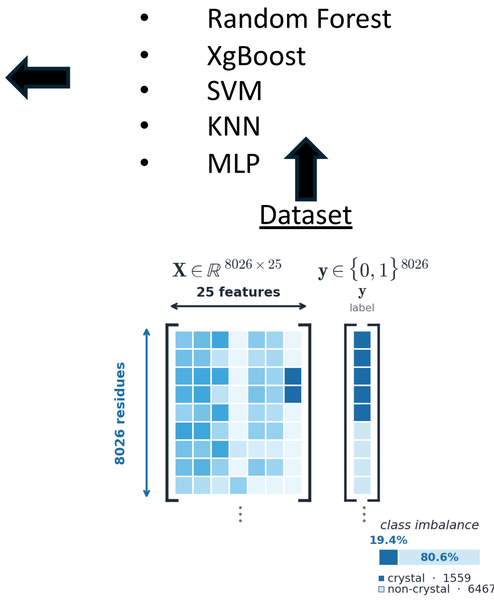
AIM

To identify the specific molecular rules governing **crystallization propensity** by evaluating key structural features through optimized machine learning algorithms, ultimately providing a predictive tool to guide the **mutation design in nanobodies**.

METHOD



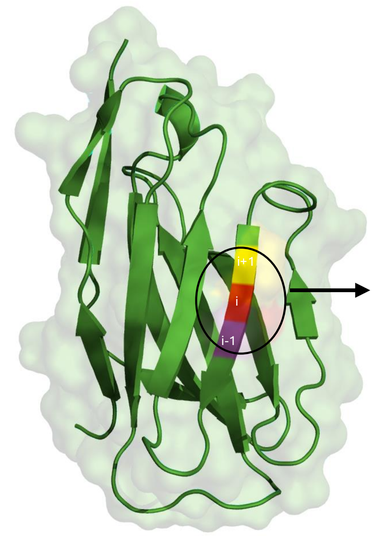




Feature engineering

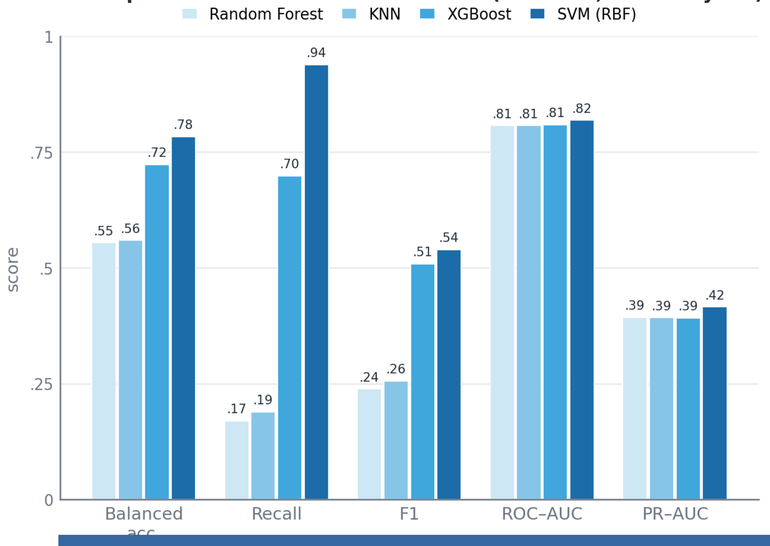
Feature extraction per residue

- Solvent accessible surface area
- Protrusion index
- Fractional exposure
- Spatial aggregation propensity
- Charge(sum of partial charges)
- Hydrophobicity
- Depth index



i_{nb} [≤ 5 Å and ≤ 7 Å
h
Type of residue any type, hydrophobicpolar-charged, hydrophobic-charged, & charged-polar

RESULTS



Model performance

held-out residues · n = 1606 · 19.4% crystal

Model	Precision	Recall	F1	Balanced acc.	ROC-AUC	PR-AUC
Random Forest	**.40**	.17	.24	.55	.81	.39
KNN	.40	.19	.26	.56	.81	.39
XGBoost	.40	.70	.51	.72	.81	.39
SVM (RBF)	.38	**.94**	**.54**	**.78**	**.82**	**.42**

shaded = best model for that metric

DISCUSSION

These preliminary results show comparable ranking power across all four classifiers (ROC-AUC ≈ 0.81). The imbalance-aware SVM and XGBoost recover crystal-contact residues well (recall 0.94 and 0.70), suggesting the SAP-adjacent-FE and surface-exposure descriptors carry real signal, though precision stays low at 19.4% positives. Full 10-fold cross-validation and SHAP analysis are still pending.

FUTURE WORK

1. 10-fold cross-validation, structure-grouped to prevent leakage
2. SMOTE oversampling to counter class imbalance
3. Compare SMOTE against current class-weighting approach
4. In silico mutations
5. Finally,validate these mutations *in vitro*

REFERENCES

References: 1. Walsh, G., Walsh, E. Biopharmaceutical benchmarks 2022. Nat Biotechnol 40, 1722–1760 (2022). https://doi.org/10.1038/s41587-022-01582-x
.Chattaraj et al (2025c). Investigating structural biophysical features for antigen-binding fragment crystallization via machine learning. Molecular Systems Design & Engineering, 10(5), 377–393. https://doi.org/10.1039/d4me00187g

ACKNOWLEDGEMENTS

Supported by the PROCRYSTAL project funded by the European Union's Marie Skłodowska-Curie Actions Doctoral Networks programme under grant agreement No **101169471**.