

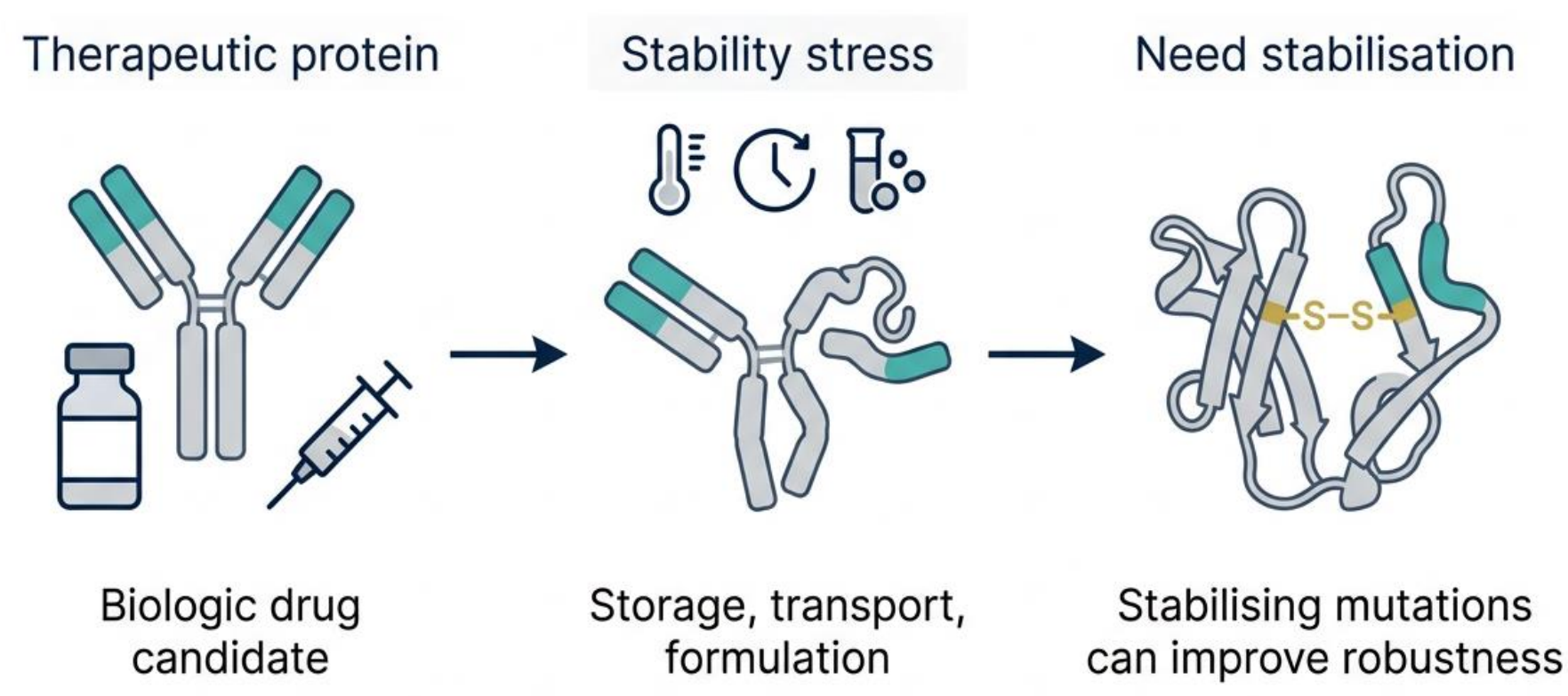
Machine Learning-Guided Protein Disulfide Bond Engineering

Binjie Chen¹, Giovanni Stracquadanio²

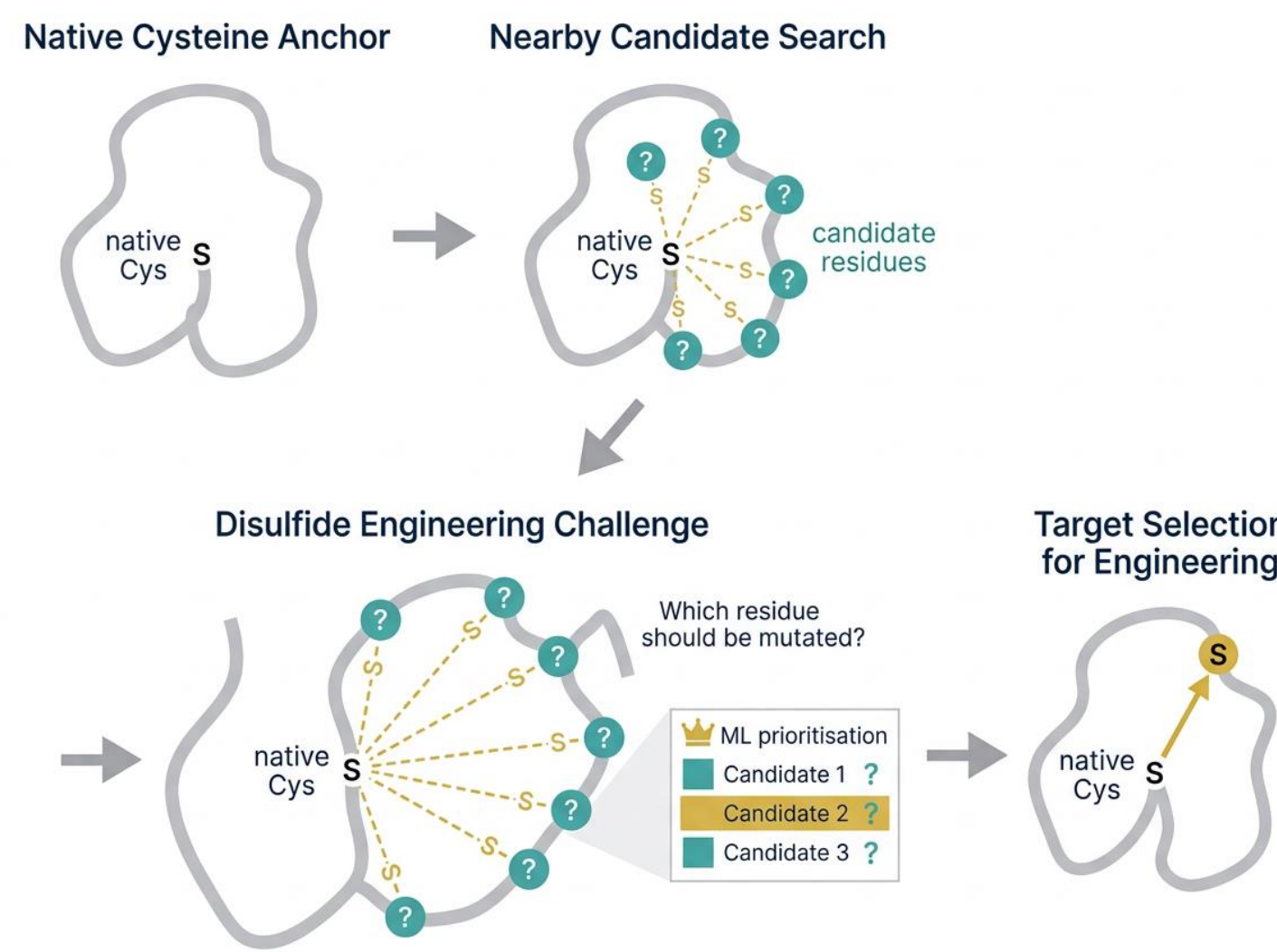
¹ School of Informatics, The University of Edinburgh, Edinburgh EH8 9AB, U.K.

² School of Biological Sciences, The University of Edinburgh, Edinburgh EH9 3BF, U.K.

Motivation

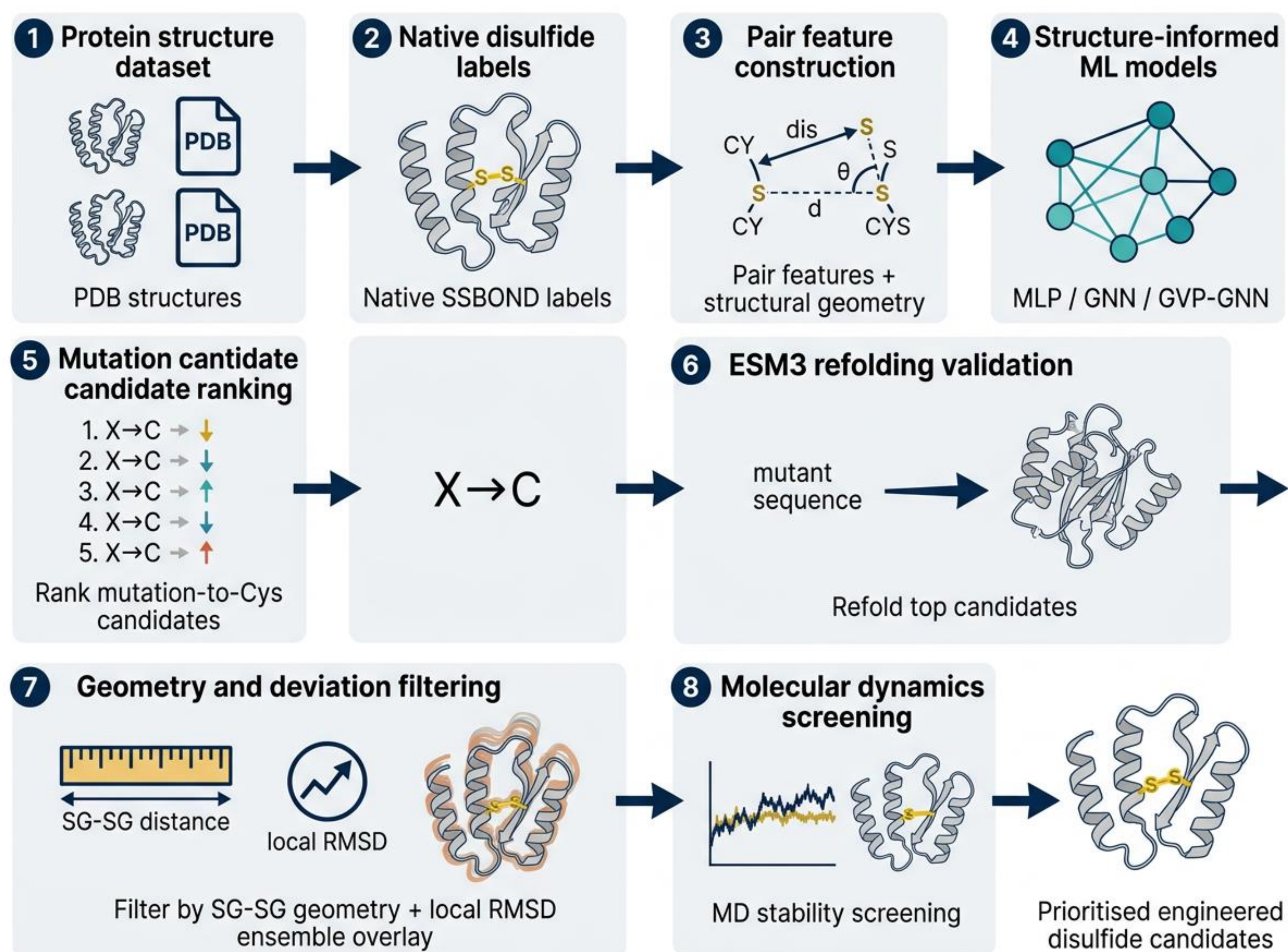


Protein stability is a key challenge for biologic drug development, motivating computational strategies for stabilising mutation design.



Protein stability is a key bottleneck in therapeutic and industrial protein engineering. Engineered disulfide bonds can improve thermal stability and robustness, but selecting mutation sites remains difficult because successful candidates must satisfy both local disulfide geometry and global structural compatibility.

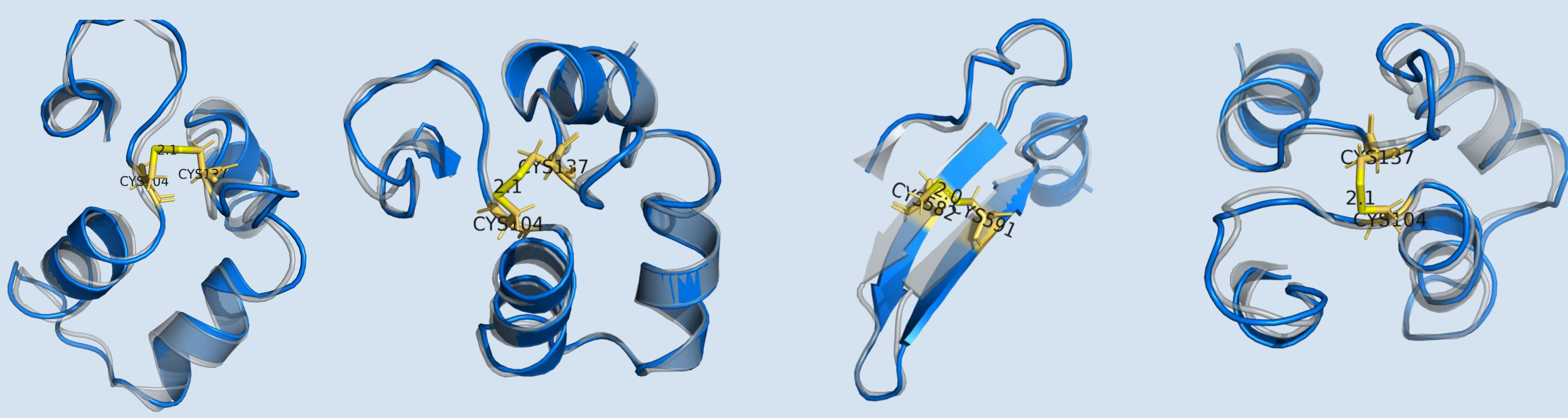
Machine learning-guided disulfide engineering pipeline



Native disulfide annotations from PDB structures are used as proxy supervision to learn disulfide-like structural patterns. The trained models rank mutation-to-cysteine candidates, which are then refolded, geometrically filtered and evaluated by molecular dynamics before experimental prioritisation.

Results visualization: top candidate bonds

Representative local views of shortlisted engineered disulfide candidates, showing candidate cysteine pairs in their surrounding backbone environment.



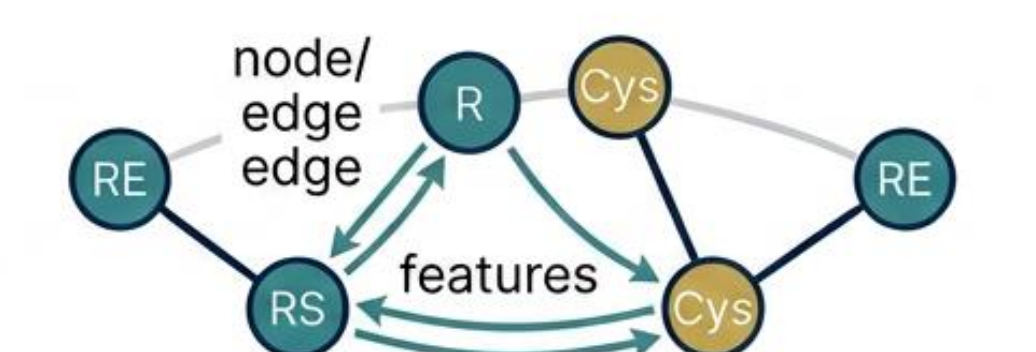
Shortlisted candidates should be evaluated after refolding and MD as per the figure illustrated on the right. Stable designs should maintain S-S distances near the disulfide-compatible range while showing limited local and global backbone deviation.

Structure-Informed Models

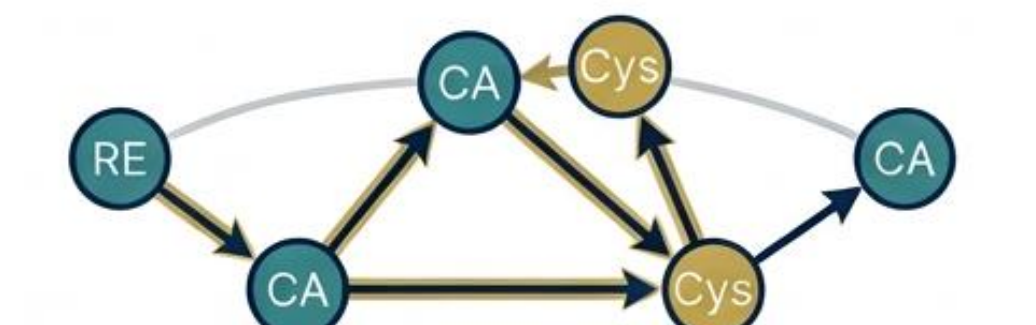
Large MLP
Pair-level structural features

distance sequence separation orientation ...

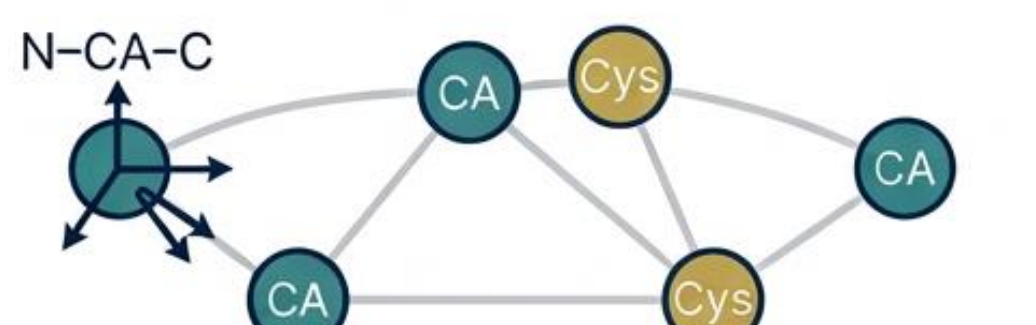
Scalar GNN
Residue graph message passing



CA-GVP
CA direction-aware graph



Backbone-GVP
Local backbone frames



Capacity-matched model comparison

Held-out PDB split; native disulfide labels used as proxy supervision

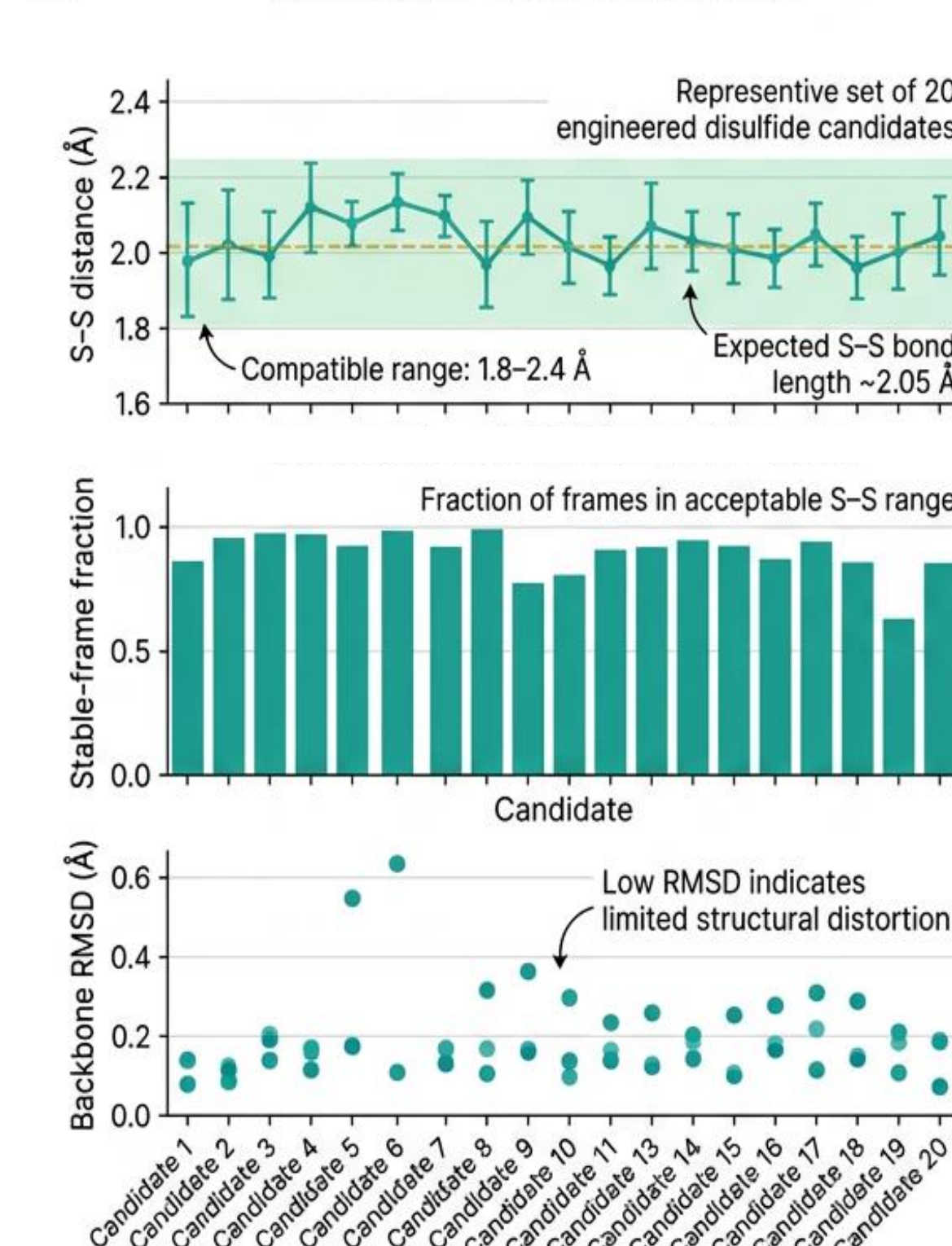
Model	Input representation	Params	Test AP	Test AUC
Large MLP	Pair features	76k	0.936	0.988
Scalar GNN	Residue graph	76k	0.928	0.986
CA-GVP	CA directions	77k	0.956	0.991
Backbone-GVP	Backbone frames	77k	0.961	0.992

AP = average precision; AUC = ROC-AUC. Parameter counts are approximately matched.

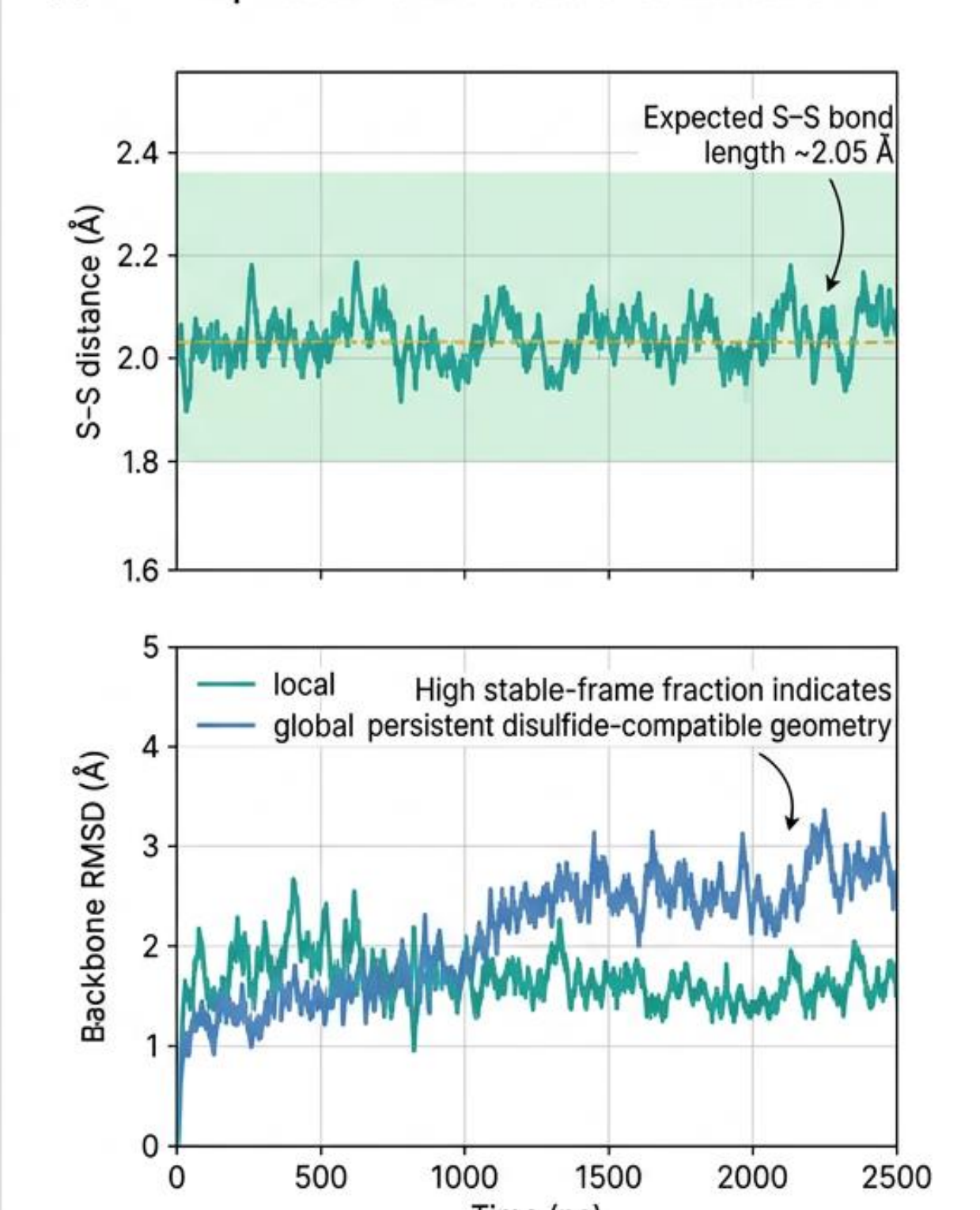
Direction-aware graph models gave the strongest held-out disulfide-pair prediction under matched capacity.

Schematic validation summary

A Candidate-level MD screen



B Representative stable candidate



Acknowledgements