Prioritization of new molecule designs using QSAR models: 2D- and 3D-QSAR studies on SARS-CoV-2 M^{pro} inhibitors

Oliver Hills, Matthew Kondal & Natercia Braz

Cresset, Cambridgeshire, UK   oliver.hills@cresset-group.com   cresset-group.com

Abstract

The viral main protease M^{pro} is a crucial enzyme for the replication of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Because of its key role, M^{pro} has received much attention as a potential target for novel antivirals.[1-6] Using a dataset of 76 M^{pro} inhibitors with known activity and a common binding mode, robust and predictive machine learning (ML) and 3D-Field Quantitative Structure Activity Relationships (QSAR) models were developed, suggesting novel design edits required to maximise potency.

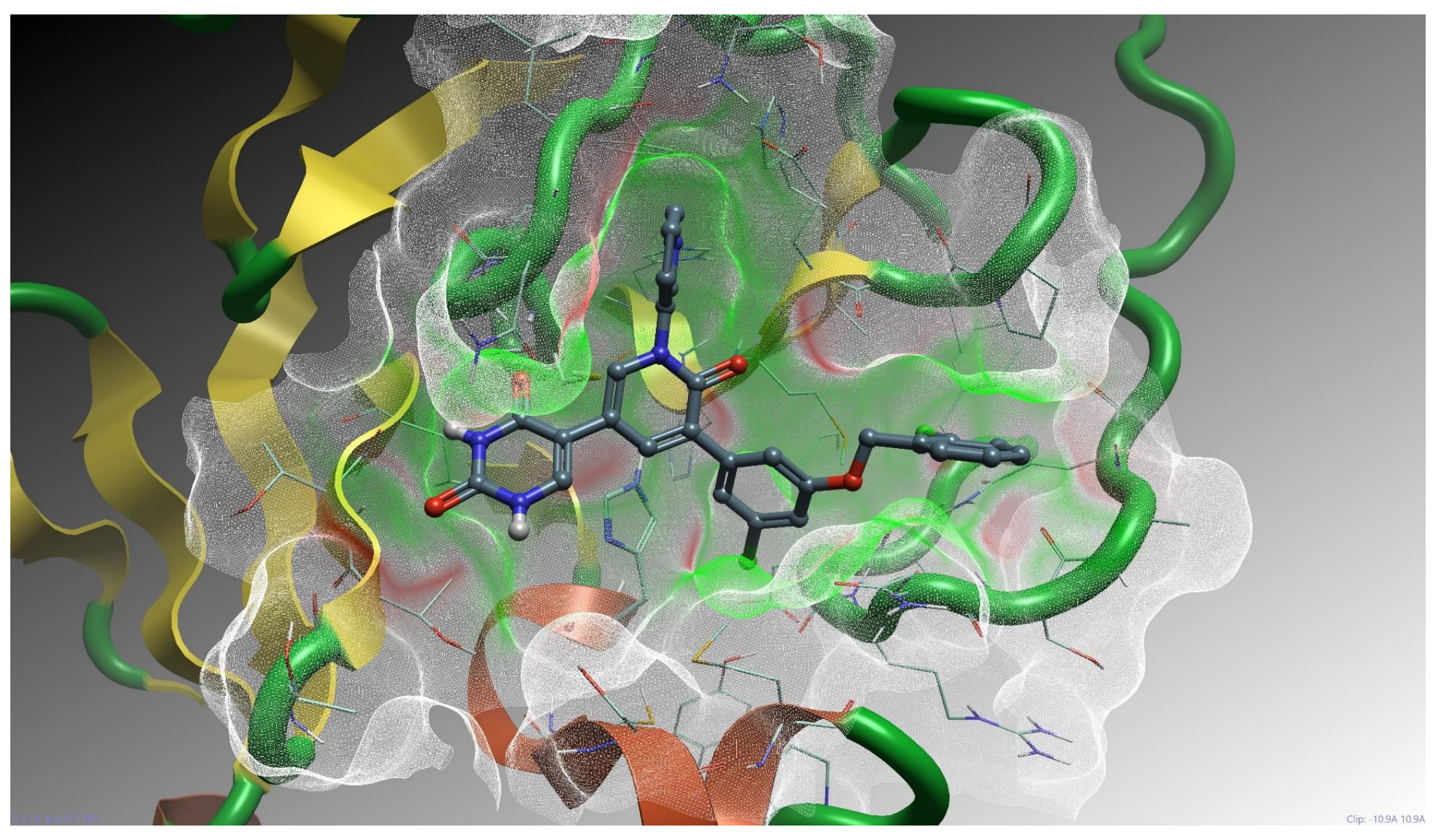


Figure 1: Crystal structure of the SARS-CoV-2 M^{pro} (PDB 7L13[1]) in complex with a non-covalent inhibitor. The Electrostatic Complementarity™ surface is displayed over the active site; green indicates an electrostatic match and red indicates an electrostatic clash.

Method

Datasets

76 non-covalent inhibitors with different chemotypes and an evenly distributed activity (pIC_{50}: 4.00 – 7.74) were partitioned into training set (56 molecules) and test set (20 molecules) using 26% activity stratification.

2D-QSAR

2D physico-chemical descriptors were computed using RDKit[7] natively within Flare™.[8] Cross-correlated descriptors were dropped by means of linear Pearson correlation matrix, producing a set of six non-redundant descriptors: MW, TPSA, #RB, NumHAcceptors, NumHDonors and RingCount. These were combined with fingerprint descriptors (RDKit, Morgan and MACCS keys) to generate 2D-QSAR regression models using supervised machine learning methods: Support Vector Machine (SVM), Gaussian Process Regression (GPR), Random Forest (RF), Multilayer Perceptron (MLP) and Consensus.

3D-QSAR

High-quality alignments created by Flare, particularly those based on the maximum common substructure (MCS) algorithm, generated meaningful molecular alignments with a low degree of noise (**Figure 2**). The compounds were aligned by MCS to the co-crystallized ligands of the PDB IDs 7L13[1], 7L14[1], 7QBB[5] and 8SXR[6], which were used as references (weighted average contribution) and using the 7L13 protein as an excluded volume. Alongside the above machine learning methods, 3D-QSAR regression models were generated using the Cresset Field 3D-QSAR method.

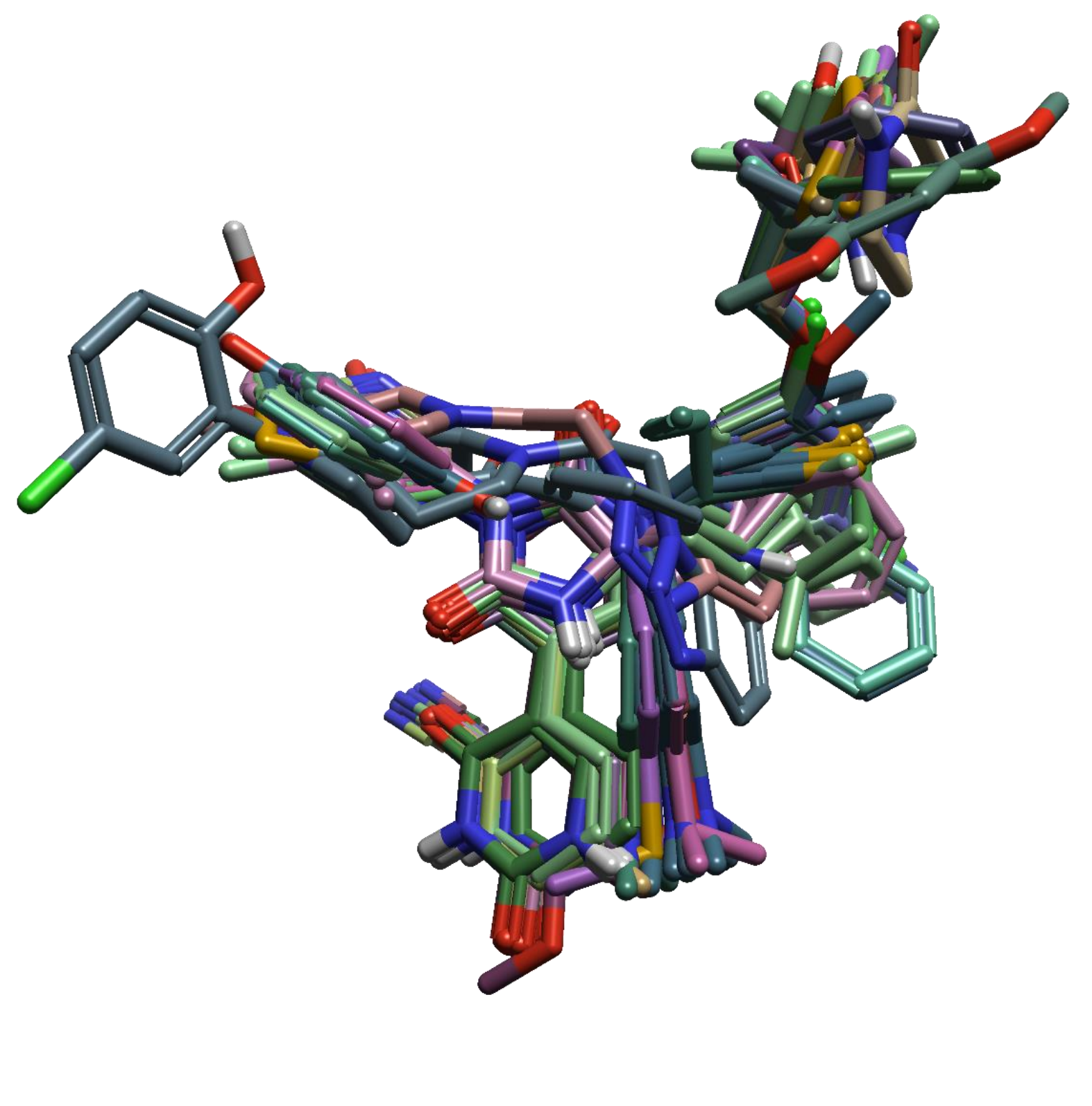


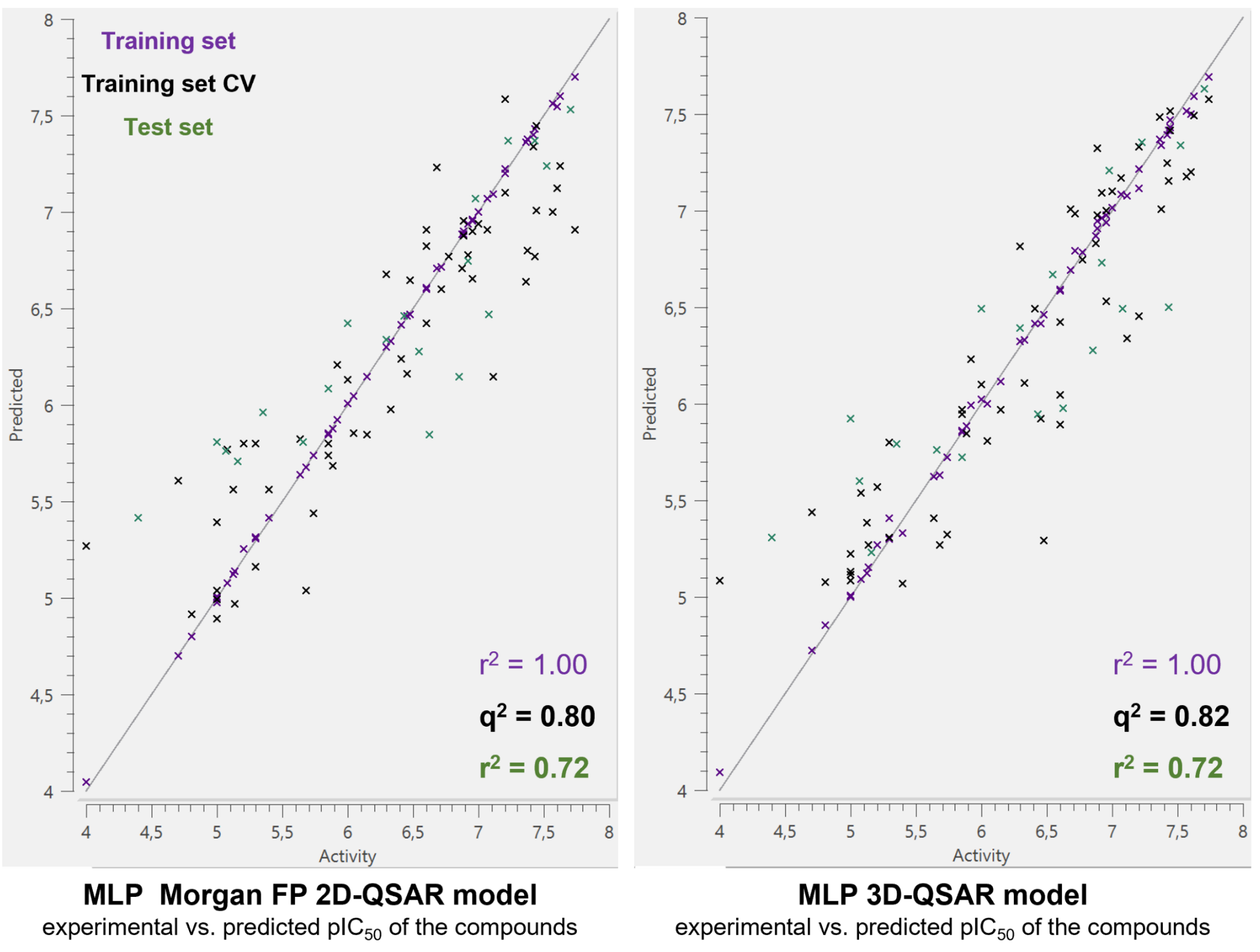
Figure 2: The dataset of 76 compounds aligned in 3D space by MCS.

Statistical Analysis

- The confidence of the generated models is high and comparable (**Table 1, Figure 2**).
- Morgan FP MLP 2D-QSAR and the MLP 3D-QSAR models are the most accurate (r^2 = 0.72).
- All these models are expected to provide the same level of accuracy in predicting the activity of new compounds.
- The good agreement between the 2D and 3D models suggests that the compounds of this dataset act via a similar mechanism.
- RDKit 2D descriptors and fingerprints are good alternatives to Cresset 3D descriptors for building predictive ML models.
- The Cresset Field 3D-QSAR model coefficients identify functionality about the molecular frame critical for potency.

Table 1: Comparison of the different QSAR models measured and predicted statistics

QSAR type	Regression model	r^2 training set	q^2 training set CV	r^2 test set
2D-QSAR (6 physico-chemical descriptors)	MLP	0.91	0.68	0.69
	GPR	0.89	0.73	0.67
	Consensus	0.89	0.74	0.65
	RF	0.86	0.74	0.62
	SVM	0.86	0.75	0.61
2D-QSAR (fingerprints (FP))	MLP (Morgan FP)	1.00	0.80	0.72
	SVM (RDKit FP)	1.00	0.83	0.63
	SVM (MACCS keys)	0.96	0.80	0.50
3D-QSAR	MLP	1.00	0.82	0.72
	Field QSAR	0.96	0.81	0.71
	Consensus	0.99	0.82	0.70
	SVM	0.98	0.82	0.70
	GPR	0.99	0.77	0.70
	RF	0.99	0.82	0.70



MLP Morgan FP 2D-QSAR model
experimental vs. predicted pIC_{50} of the compounds

MLP 3D-QSAR model
experimental vs. predicted pIC_{50} of the compounds

Figure 3: MLP Morgan FP 2D-QSAR (left) and 3D-QSAR (right) models. Experimental *vs.* predicted activity of the compounds in the training set (purple), training set Cross Validation (black) and the test set (green).

References

1. Chun-Hui Zhang, et al., *ACS Cent. Sci.* **2021**, 7, 467–475, https://doi.org/10.1021/acscentsci.1c00039
2. Chun-Hui Zhang, et al., *ACS Med. Chem. Lett.* **2021**, 12, 1325–1332, https://doi.org/10.1021/acsmedchemlett.1c00326
3. Maya G. Deshmukh, et al., *Structure* **2021**, 29, 823–833, https://doi.org/10.1016/j.str.2021.06.002
4. William L. Jorgensen, Patent WO 2022/150584 A1
5. Andreas Luttens et. al., *J. Am. Chem. Soc.* **2022**, 144, 2905–2920, https://doi.org/10.1021/jacs.1c08402
6. Jimena Perez-Vargas et. al., *Emerg. Microbes Infect.* **2023**, 12, 2246594, doi.10.1080/22221751.2023.2246594
7. RDKit: Open-source cheminformatics. https://www.rdkit.org
8. Flare™, Cresset®, Litlington, Cambridgeshire, UK; https://www.cresset-group.com/software/flare/; Cheeseright T., Mackey M., Rose S., Vinter, A.; Molecular Field Extrema as Descriptors of Biological Activity: Definition and Validation *J. Chem. Inf. Model.* **2006**, 46 (2), 665-676.

Field 3D-QSAR Model Visualization and Interpretation

The Cresset Field 3D-QSAR method offers the advantage over ML methods, in that the visual inspection of the model coefficients identifies regions where the model predicts strong effects on activity.

Figure 4 illustrates the electrostatic and steric model coefficients superposed to the most potent molecule (**37**, pIC_{50} = 7.74). Regions of favorable negative electrostatic coefficients are observed in the amide-carbonyl of the core ring and the nitrogen atom of the pyridine unit, which implies that a less positive charge on these regions improves activity. Additionally, the large green dots point out regions of favorable steric coefficients near the 2-chlorobenzyl moiety, which in combination to the high steric variance verified this is the best moiety to model to increase potency.

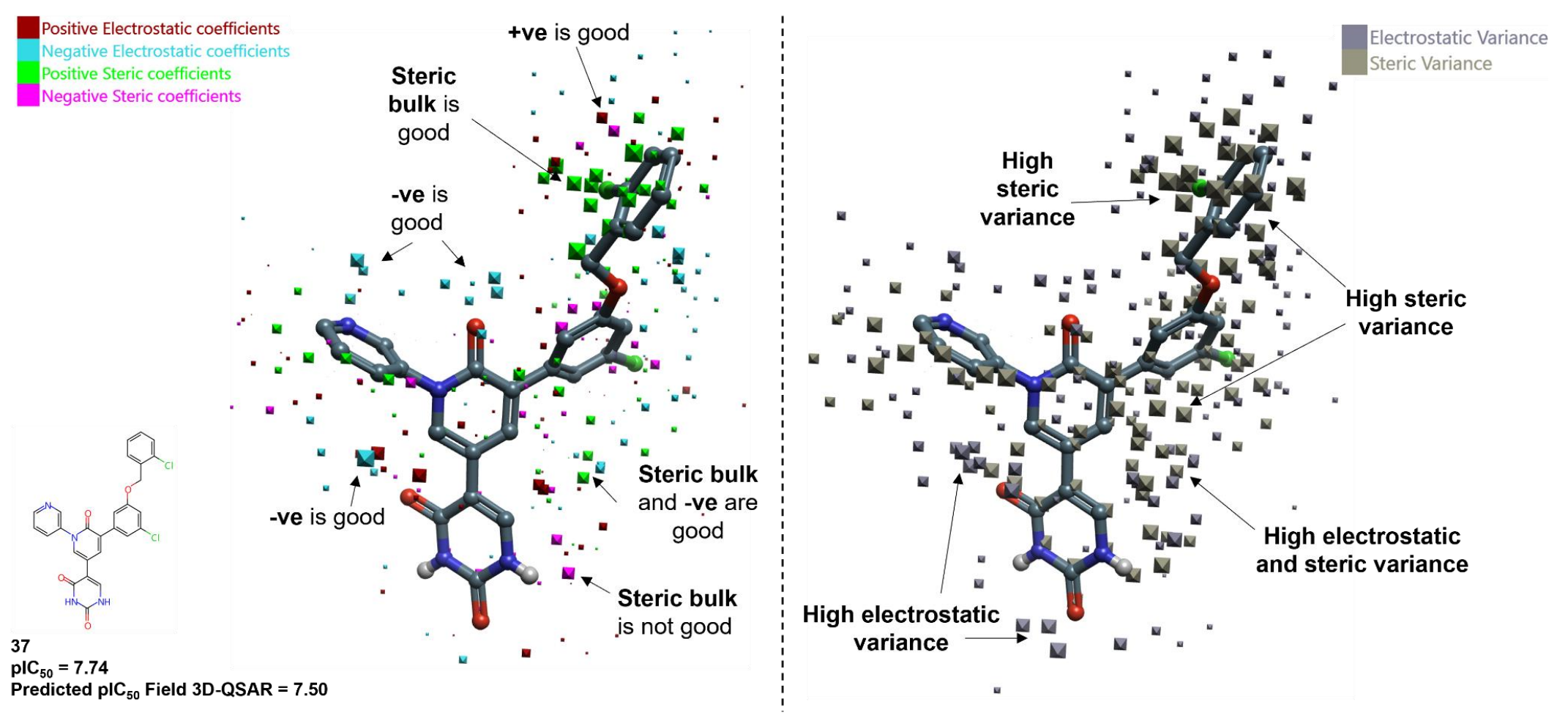

🟥 Positive Electrostatic coefficients
🟦 Negative Electrostatic coefficients
🟩 Positive Steric coefficients
🟪 Negative Steric coefficients
+ve is good
Steric bulk is good
-ve is good
-ve is good
Steric bulk and -ve are good
Steric bulk is not good
37
pIC_{50} = 7.74
Predicted pIC_{50} Field 3D-QSAR = 7.50
Electrostatic Variance
Steric Variance
High steric variance
High steric variance
High electrostatic and steric variance
High electrostatic variance

Figure 4: Model coefficients for the M^{pro} Field QSAR model. Electrostatic and steric coefficients (left); electrostatic and steric variance (right), using the most potent molecule (**37**) as reference. Compound numbering is according to the patent WO2022/150584A1.[4]

Furthermore, the relevance of the 2-chlorobenzyl alcohol group is highlighted by comparing the field contributions of compound **37** with similar molecules (**Figure 5**).

- The absence of this group in compound **8** has an unfavorable electrostatic contribution that decreases activity by *ca*. 2.5 log units.
- Large and unfavorable electrostatic and steric contributions are observed with the substitution of the aromatic ring, causing a decrease in activity of *ca*.1 log unit.
- The presence of a hydroxyl group such as in compound **28** has a strong unfavorable electrostatic contribution which decreases its predicted activity. **28** does present a clear favourable steric contribution that rationalizes its superior activity over compound **8.**

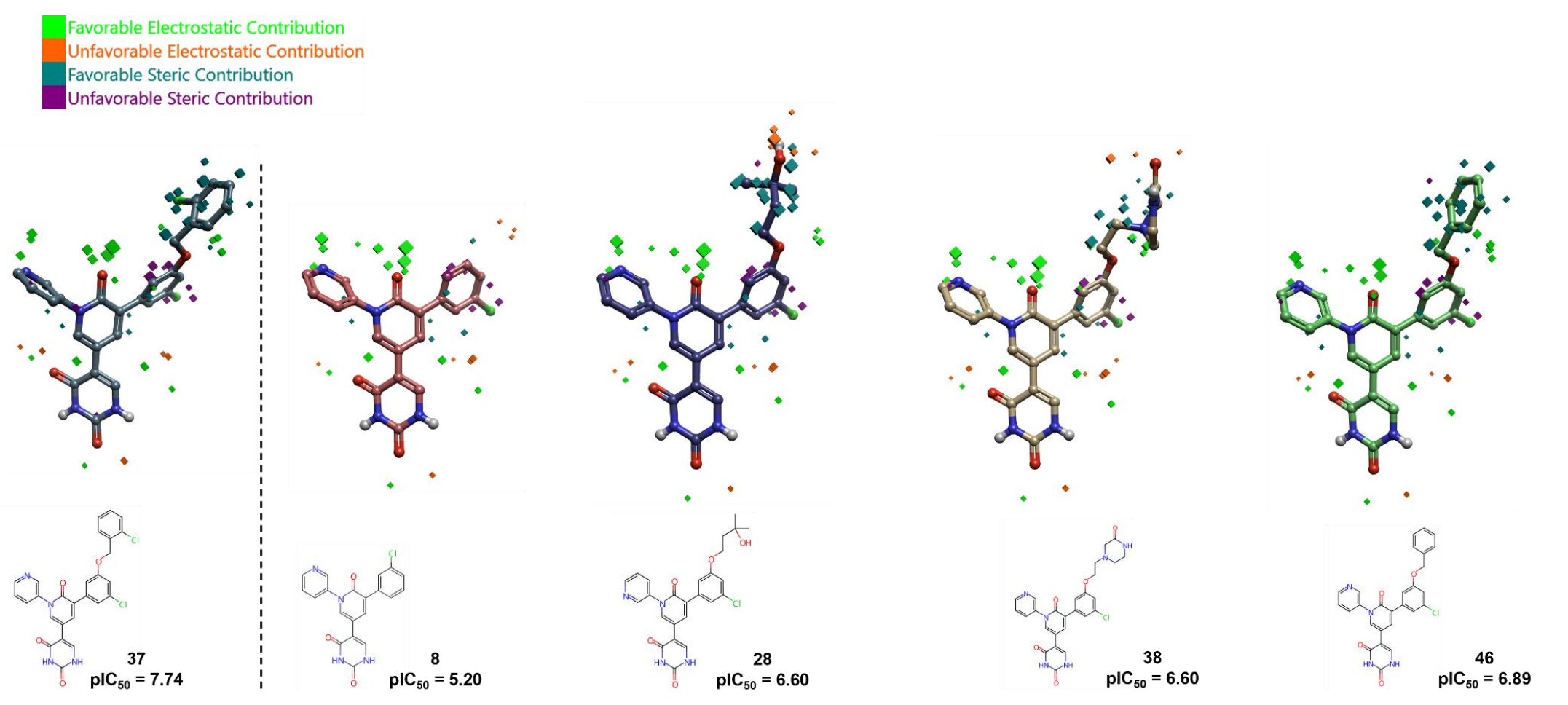

Favorable Electrostatic Contribution
Unfavorable Electrostatic Contribution
Favorable Steric Contribution
Unfavorable Steric Contribution
37
pIC_{50} = 7.74
8
pIC_{50} = 5.20
28
pIC_{50} = 6.60
38
pIC_{50} = 6.60
46
pIC_{50} = 6.89

Figure 5: SARS-CoV-2 M^{pro} 3D-QSAR field contributions to predicted activity for compounds **37**, **8**, **28**, **38** and **46**.

Conclusions

- Robust 2D-QSAR and 3D-QSAR regression models described and predicted the activity of a library of non-covalent SARS-CoV-2 M^{pro} inhibitors.
- Superior performance of the Field 3D-QSAR over the machine learning models.
- The analysis of the electrostatic and steric Field 3D-QSAR coefficients further rationalized inhibitor potency.