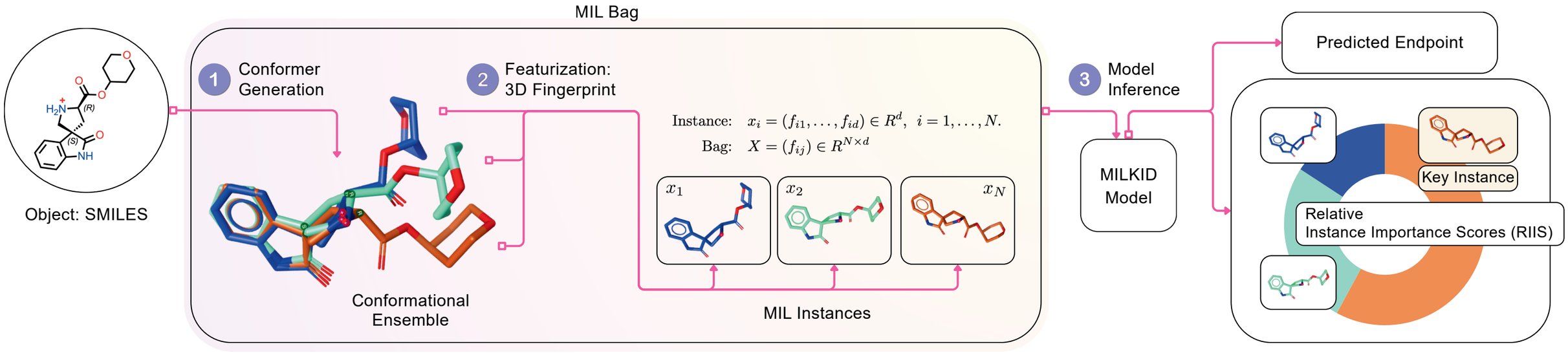
Explainable Multi-Instance Learning

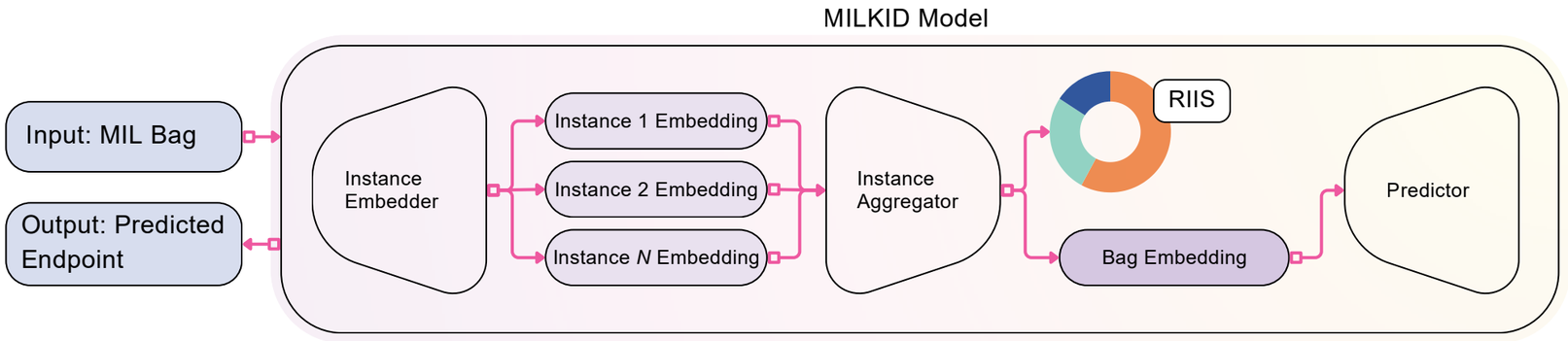
for decoding stereo-dependent biological effects

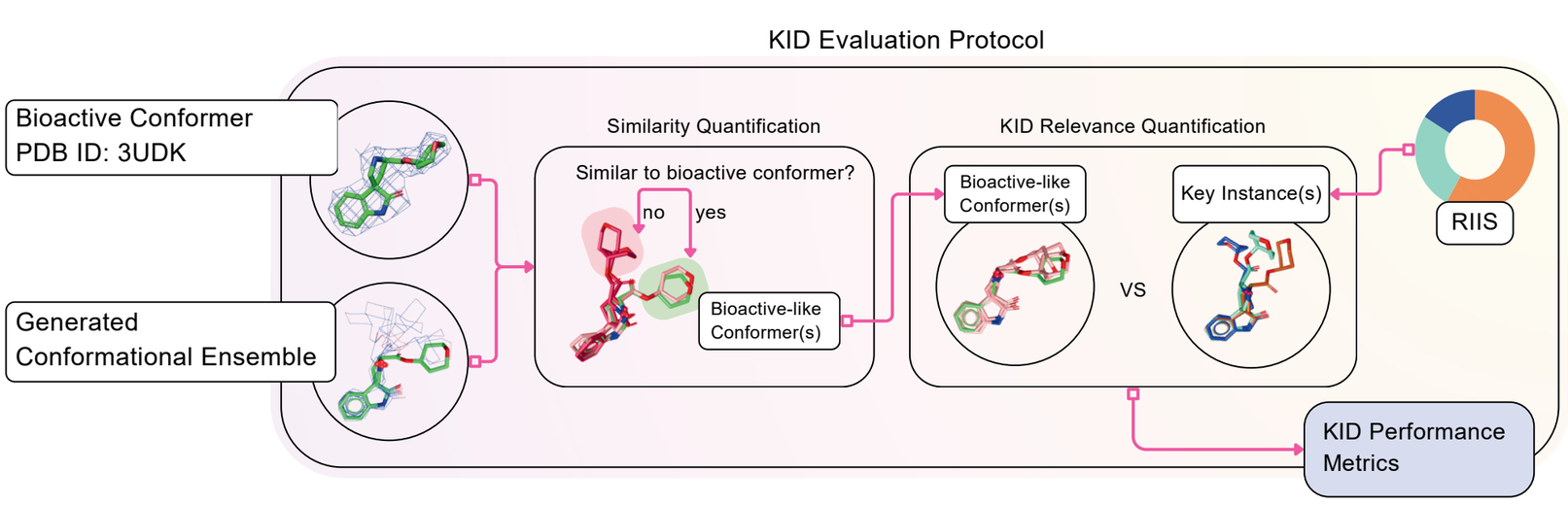
Vasilii Fastovskii[2], Gilles Marcou[1], Marc Bianciotto[2], Christoph Grebner[3], and Alexandre Varnek[1]

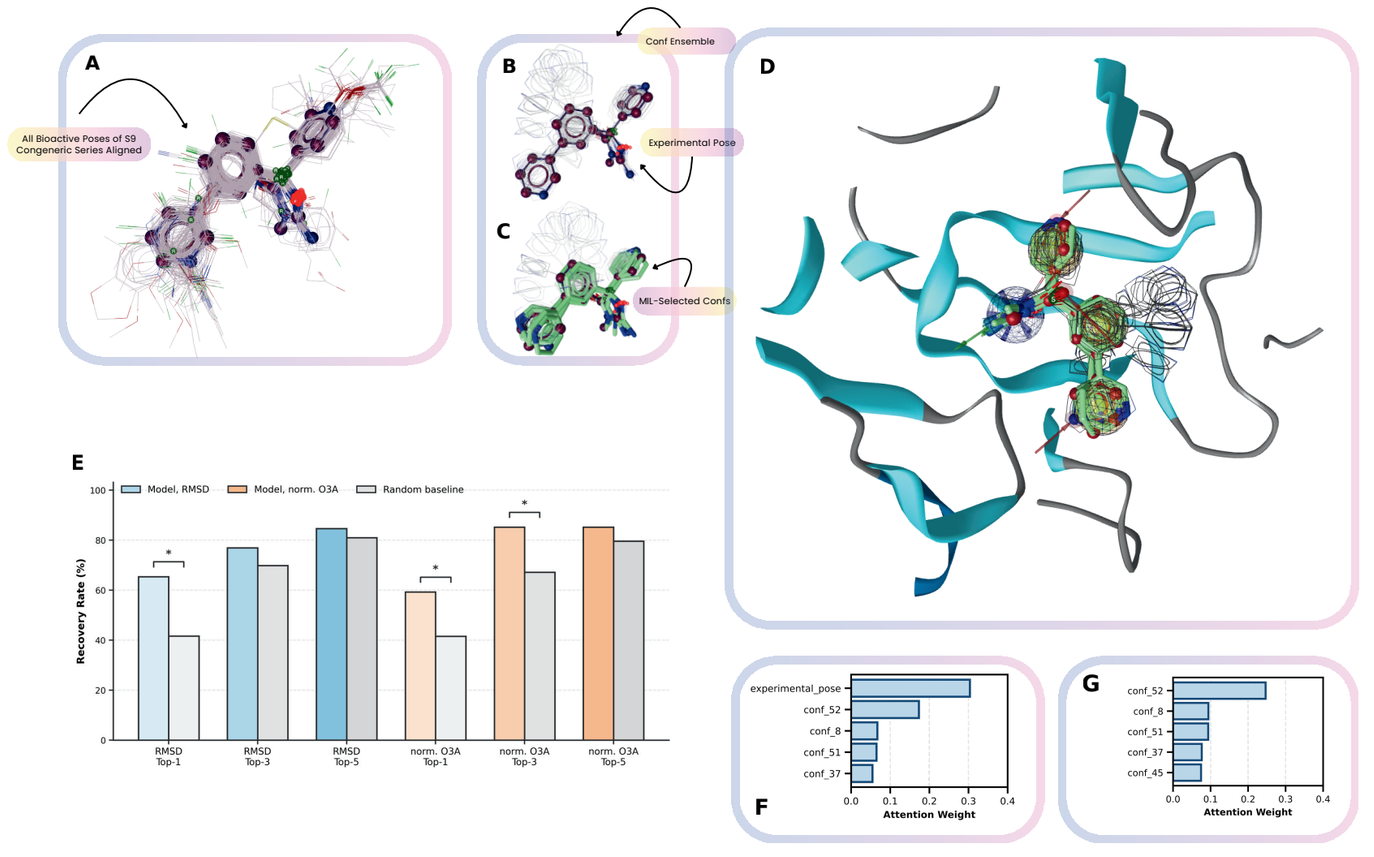
[1]Laboratory of Chemoinformatics, University of Strasbourg, Strasbourg, 67081, France
[2]Molecular Design Sciences-Integrated Drug Discovery, R&D, Sanofi, 94400 Vitry-sur-Seine, France
[3]Synthetic Molecular Design-Integrated Drug Discovery, R&D, Sanofi, 65929 Frankfurt-am-Main, Germany

Research Objective: Is it possible to develop a 3D-QSAR model that is predictive and capable of identifying bioactive molecular conformers? In this study, we introduce a multi-instance learning (MIL) workflow to address the challenge of detecting bioactive conformers for targets lacking experimental structural data through key instance detection (KID).









Interpretation of KID results in the context of Receptor Independent (RI) 3D-QSAR. (A) Bioactive conformers aligned to the reference crystal structure 2QP8; the co-crystallized ligand from 3IN3 is highlighted in red (ball-and-stick representation). (B) Conformational ensemble of 3IN3, with the experimental (bioactive) pose shown in ball-and-stick representation. (C) Most highly attended conformers of 3IN3 identified by the MIL model. (E) KID performance (D) Conformational ensemble of 3IN3 aligned to a structure-based pharmacophore model. (F–G) Comparison of the most attended conformers for two bags corresponding to the same molecule, in the presence (F) and absence (G) of a known bioactive pose.

The overall MILKID inference workflow: each molecule is represented as a bag of conformers, that is, its conformational ensemble. Each conformer is encoded using 3D roto-translation invariant descriptors. The MIL model takes the resulting bag as input and outputs (i) a molecule-level prediction and (ii) conformer-level relative importance scores reflecting the influence of each conformer on the prediction. The model-derived conformer weighting is then analyzed in a post hoc KID evaluation protocol.

We demonstrated this framework on a structurally diverse set of more than 200 BACE1 ligands with experimentally determined bioactive poses (i.e., bioactive conformers), spanning five major and eight minor congeneric series. This series-level diversity represents a key extension of our findings over previous studies.

Conclusions & Outlook

We developed a predictive MIL model capable of assigning attention weights to individual conformers. In the validation set, the bioactive-like conformer received the highest attention score for 60% of molecules. We further observed that the evaluation of KID performance depends strongly on the metric used to define similarity to the bioactive pose.

Acknowledgments

The author gratefully acknowledges the support of the AiChemist project and the Laboratory of Chemoinformatics at the University of Strasbourg for providing the scientific environment, supervision, and resources necessary to complete this project.