

SUPPORTING INFORMATION

E-GuARD: Expert-Guided Augmentation for Robust Interference Compound Detection

Vincenzo Palmacci^{1,2,#}, Yasmine Nahal^{3,4,#}, Matthias Welsch^{1,2,5}, Ola Engkvist^{4,6}, Samuel Kaski^{3,7} and Johannes Kirchmair^{1,5*}

¹ Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria

² Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences (PhaNuSpo), University of Vienna, 1090 Vienna, Austria

³ Department of Computer Science, Aalto University, Espoo, Finland

⁴ Molecular AI, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden

⁵ Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department for Pharmaceutical Sciences, University of Vienna, 1090 Vienna, Austria

⁶ Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden

⁷ Department of Computer Science, University of Manchester, Manchester, United Kingdom

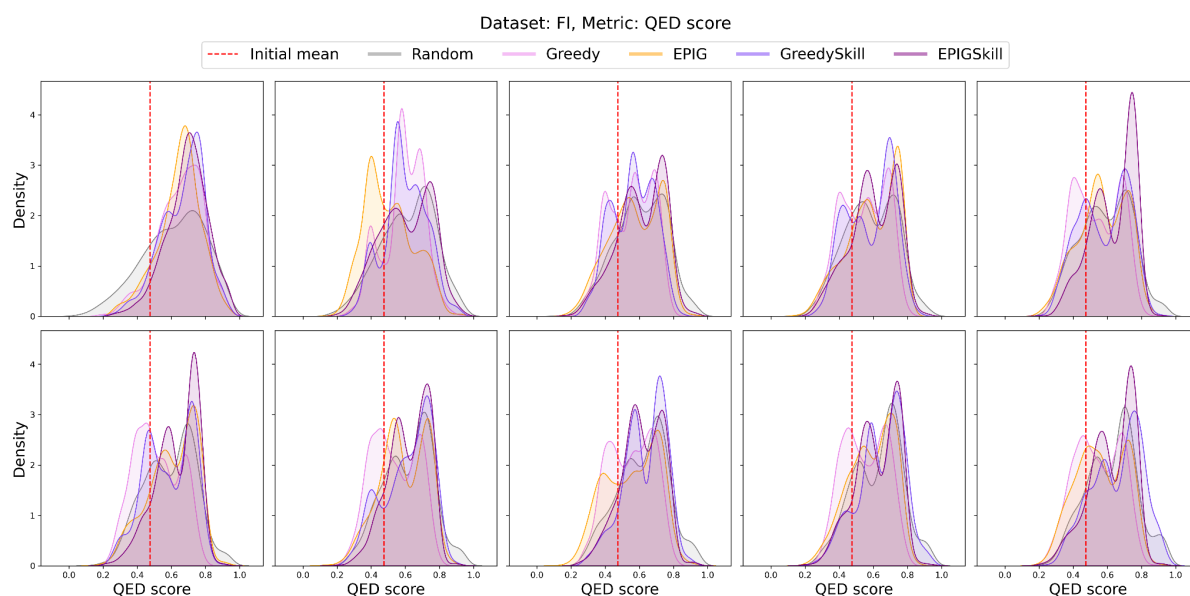
These authors contributed equally to this work.

* Corresponding author: johannes.kirchmair@univie.ac.at

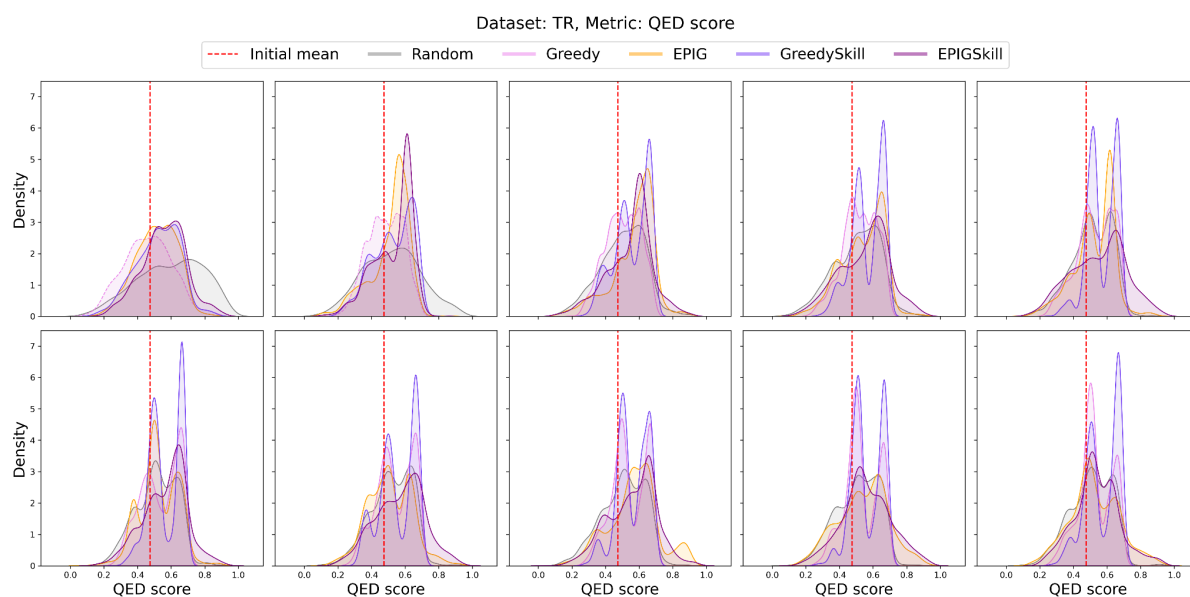
Table S1: Comparison of balanced accuracy between the Liability Predictor and E-GuARD QSIR models for four interference mechanisms using 5-fold cross-validation. For the Liability Predictor, averages, as found in the original publication, are reported. For E-GuARD, both averages and standard deviations are provided.

	FI	NI	TR	RR
Liability predictor	0.78	0.75	0.70	0.62
E-GuARD	0.81 \pm 0.01	0.71 \pm 0.02	0.73 \pm 0.01	0.66 \pm 0.02

(a)



(b)



(c)

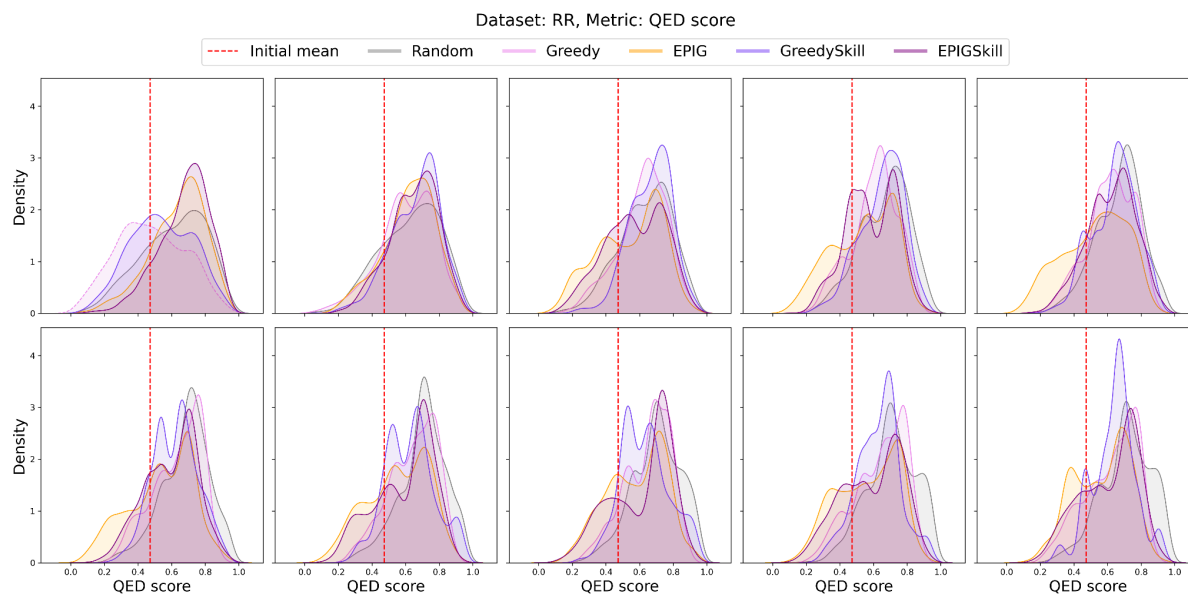


Figure S1: Distributions of QED scores of the putative interfering compounds computed across ten iterations of E-GuARD for **(a)** FI, **(b)** TR and **(c)** RR data sets. The red dashed, vertical line in each panel corresponds to the mean QED score of the interfering compounds in the initial predictor training set. Expert-guided acquisitions such as GreedySkill (in blue) and EpiGSkill (in purple) lead to higher QED scores than other acquisition strategies and the initial mean QED value, indicating that more interfering compounds possessing drug-like properties are being added to the training set throughout the E-GuARD process.