Learning Protein-Ligand Binding Affinity Using Atomic Environment Vectors

Rocco Meli,[1] Andrew Anighoro,[2] Mike Bodkin,[2] Garrett M. Morris,[3] Philip C. Biggin[1]

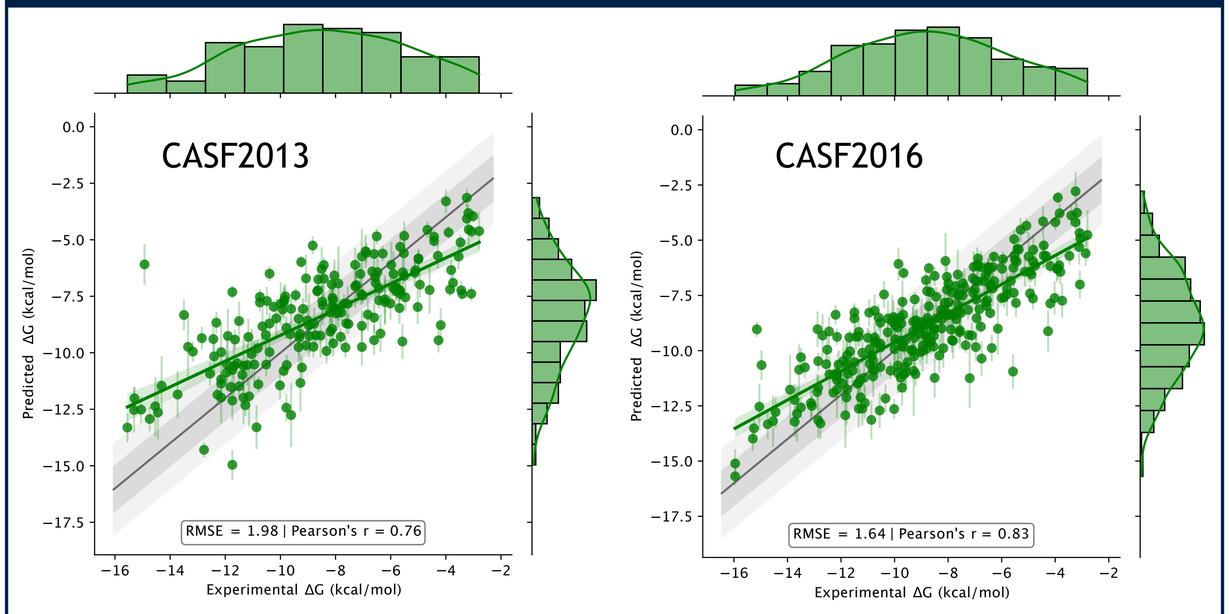
[1] Department of Biochemistry, University of Oxford [2] Evotec (UK), Abingdon [3] Department of Statistics, University of Oxford

Introduction

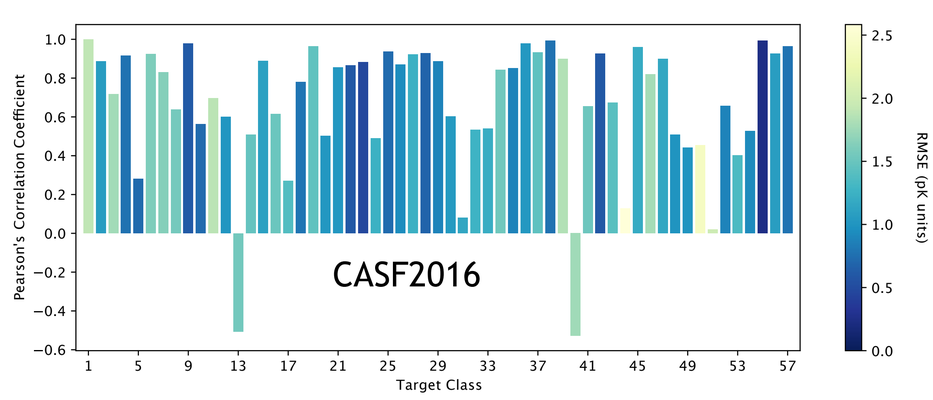
- Molecular docking is extremely useful for virtual screening
- Classical docking scoring functions (SFs) poorly predict protein-ligand (P-L) binding affinity
- Machine Learning (ML) and Deep Learning (DL) SFs outperform classical SFs
  - Convolutional neural network (CNN) SFs perform well [2]
  - CNN SFs are not rotationally invariant
  - CNN SFs need data augmentation (slow/expensive)

Can we use representations and architectures from ML/DL in quantum chemistry to develop docking scoring functions that are rotationally and translationally invariant without the need of extensive data augmentation?

Scoring:CASF2013 and CASF2016 Benchmarks

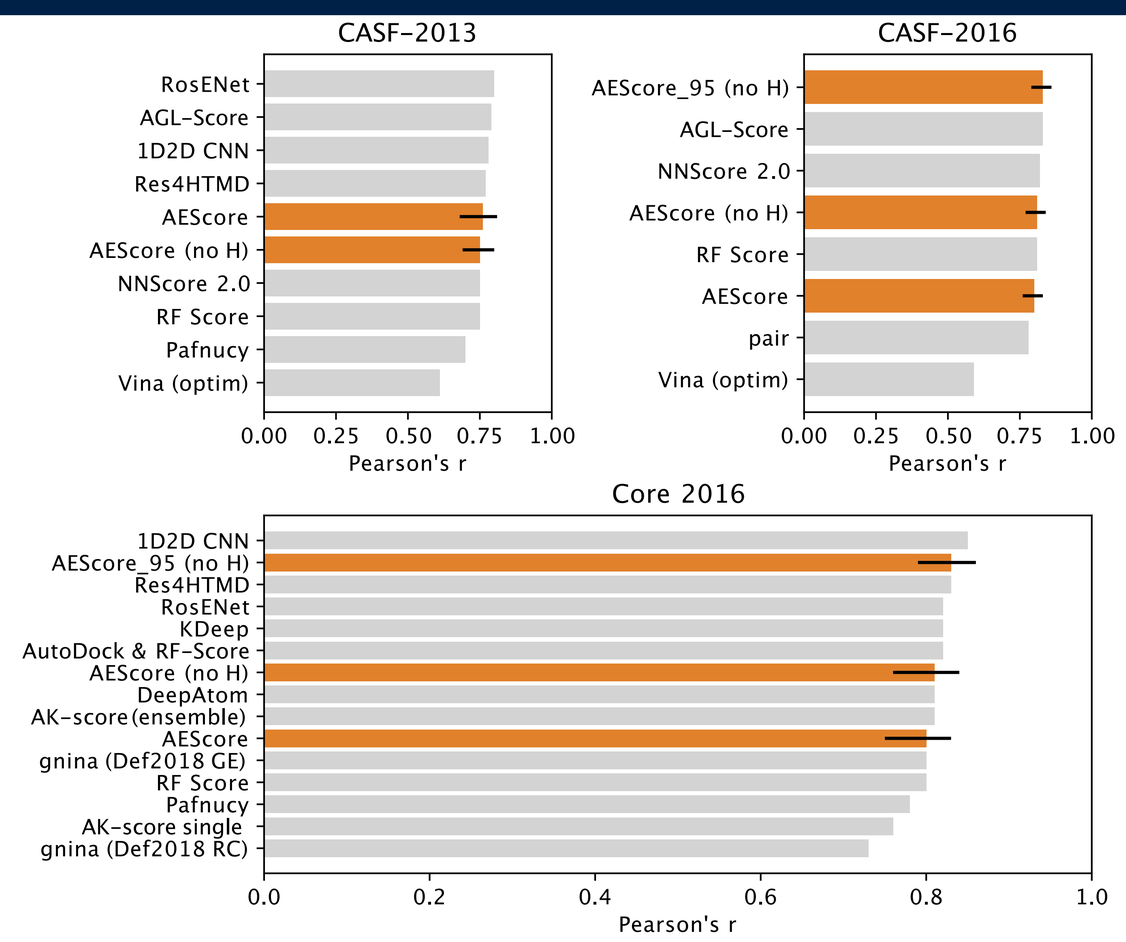


AEScore-predicted ΔGs show good correlation with experimental ΔGs



Per-class variability leaves space for fine tuning (transfer learning)

Scoring: Comparison with Other Deep Learning Scoring Functions



AEScore performs competitively with other state-of-the-art (SOTA) deep learning SFs.

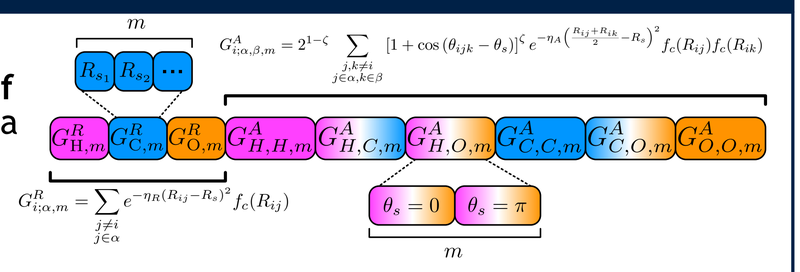
References

[1] Meli *et al.*, J.Cheminform. **13**, 59 (2021)
[2] McNutt *et al.*, J. Cheminform. **13**, 43 (2021)
[3] Behler and Parrinello, PRL **98**, 146401 (2007)
[4] Smith *et al.*, Chem. Sci. **8**, 3192 (2017)
[5] Gao *et al.*, J. Chem. Inf. Model., 60, 3408 (2020)
[6] Gomes *et al.*, arXiv, 703.10603 [cs.LG] (2016)

Atomic Environment Vectors (AEVs) and NNs [3,4,5]

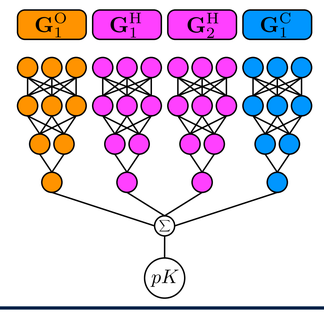
AEVs capture the local chemical environment of an atom; they represent a fine-grained and environment-dependent atom type (fingerprint). Locality is enforced by a cutoff function:

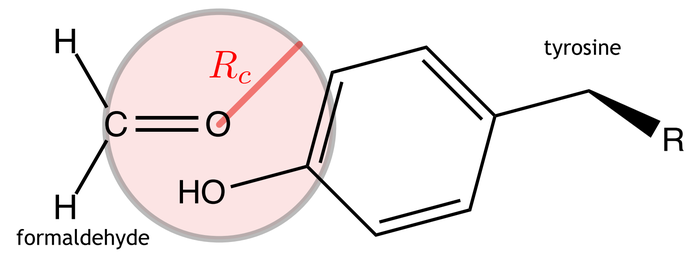
$$f_c(R_{ij}) = \begin{cases} \frac{1}{2}\left[\cos\left(\frac{\pi R_{ij}}{R_c}\right)+1\right] & R_{ij} \le R_c \\ 0 & R_{ij} > R_c \end{cases}$$



The AEV for oxygen in formaldehyde interacting with tyrosine encodes the radial and angular distribution of all atoms within the cutoff

By construction, **AEVs are rotationally and translationally invariant.**





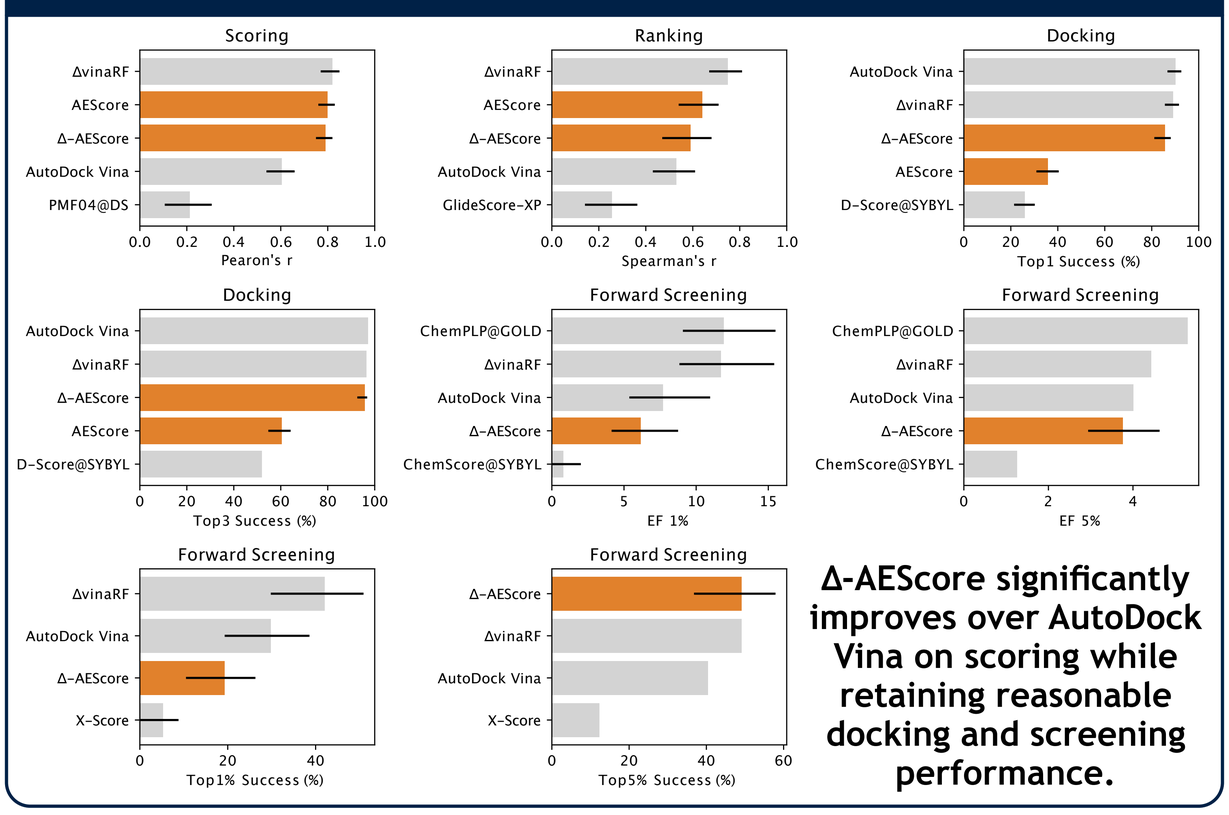
A collection of NNs allows to compute atomic contributions from AEVs, which are summed together to obtain the final prediction. This architecture has been very successful to build NN potentials [3,4].

Δ-learning

Can the method also be used for docking and screening? We can learn corrections Δ to AutoDock Vina instead of the binding affinity directly:

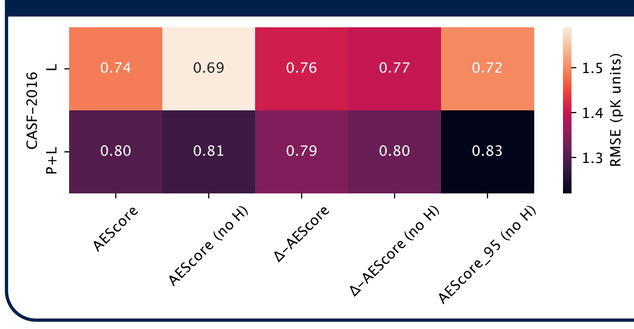
$$\Delta\text{-AEScore} = S_{\text{Vina}} + \Delta$$

Scoring, Ranking, Docking and Screening (CASF2016)



Δ-AEScore significantly improves over AutoDock Vina on scoring while retaining reasonable docking and screening performance.

Ligand-Based Performance



The model extracts a lot of information from the ligand. Introducing a full thermodynamic cycle could improve the model [6]:

$$\Delta G = G_{\text{PL}} - G_{\text{P}} - G_{\text{L}}$$

Conclusions

- SOTA performance on binding affinity prediction (AEScore)
- Rotationally and translationally invariant predictions (by construction)
- Faster training than CNN SFs (no data augmentation required)
- Δ-AEScore retains performance of AESCore but improves docking/screening
- Many places for further improvements (architecture, AEVs, data, ...)

- **Architectures used in quantum ML (QML) can be used to develop SFs**

BBSRC bioscience for the future
OXFORD Interdisciplinary Bioscience DTP
Department of Biochemistry
evotec

@RMeli
@RoccoMeli