LIE

Comprehensive and automated Linear Interaction Energy based binding-affinity prediction
Applied to Cytochrome P450 aromatase inhibitors

Marc van Dijk   Luigi Capoferri   Nico P.E. Vermeulen   Daan P. Geerke (d.p.geerke@vu.nl)

Why binding affinity prediction

Protein-ligand binding free energy prediction extends the 'docking problem' with the requirement of an accurate understanding of the chemical space leading to interactions and their relation to binding affinity. Powerful *in-silico* methods able to predict favourable interaction to the target protein and unwanted binding to off-target proteins during early-stage drug development can help to prevent failures and waste of resources during later stages.
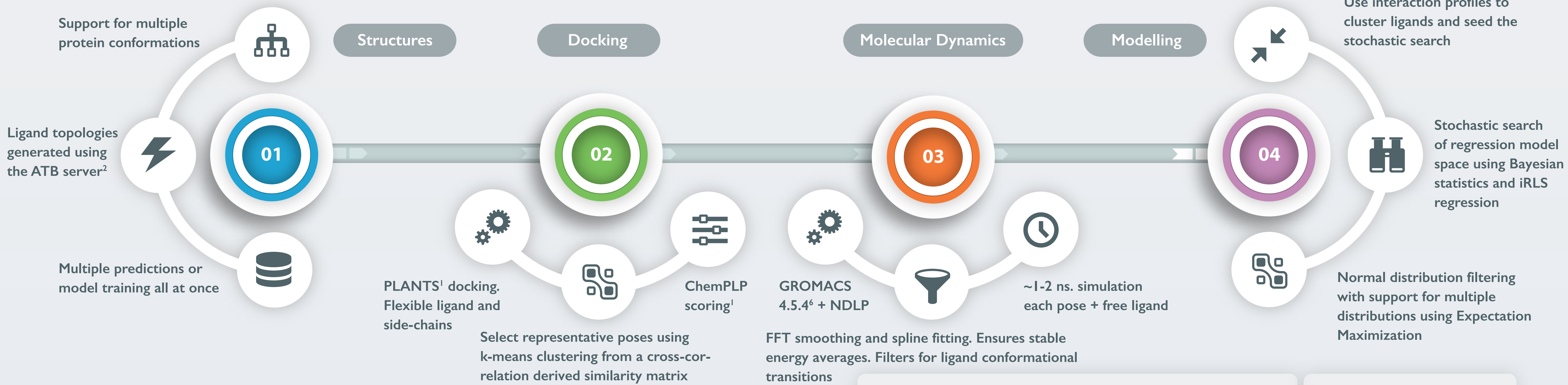
Challenges

Accurate prediction of the interaction between a ligand and a protein relies on:
• Locating one or more favourable and relevant ligand interaction poses (docking).
• Dealing with flexibilty of the protein and ligand leading to the interaction.
• Accurate description of the physicochemical interaction space (force field).
• Correlating (simulation) descriptors to binding affinity (statistical modelling).

Method LIE prediction workflow

Our approach to these challenges relies on docking to sample ligand-binding conformations followed by short MD to retrieve binding free energies. Interaction dynamics is further accounted for by using multiple ligand binding poses in a Boltzmann-weighting scheme. Binding affinity regression models are trained using a semi-autonomous learning method introduced here ● that uses iterative Linear Interaction Energy (iLIE) thereom[3,4].

Support for multiple protein conformations

Ligand topologies generated using the ATB server[2]

Multiple predictions or model training all at once

Structures **Docking** **Molecular Dynamics** **Modelling**

01   02   03   04

Use interaction profiles to cluster ligands and seed the stochastic search

Stochastic search of regression model space using Bayesian statistics and iRLS regression

Normal distribution filtering with support for multiple distributions using Expectation Maximization

PLANTS[1] docking. Flexible ligand and side-chains

ChemPLP scoring[1]

Select representative poses using k-means clustering from a cross-correlation derived similarity matrix

GROMACS 4.5.4[6] + NDLP

~1-2 ns. simulation each pose + free ligand

FFT smoothing and spline fitting. Ensures stable energy averages. Filters for ligand conformational transitions

● LIE equation with support for multiple ligand poses

$$\Delta G_{calc} = \alpha \sum_{i}^{N} W_i \left(\left\langle V_{lig-surr}^{vdW} \right\rangle_{bound,i} - \left\langle V_{lig-surr}^{vdW} \right\rangle_{free} \right) + \beta \sum_{i}^{N} W_i \left(\left\langle V_{lig-surr}^{el} \right\rangle_{bound,i} - \left\langle V_{lig-surr}^{el} \right\rangle_{free} \right) + \gamma$$

● Boltzmann weighting function for multi-pose LIE equation[3]

$$W_i = \frac{e^{-\Delta G_{calc,i}/k_B T}}{\sum_{i}^{N} e^{-\Delta G_{calc,j}/k_B T}}$$

Results LIE workflow applied to CYP19A1 aromatase

The LIE workflow was applied to a set of 132 structurally diverse aromatase (CYP19A1) inhibitors with known binding affinties derived from an industrial screening study. The high structural diversity in the dataset makes it unlikely for one model to explain the full dataset. Our semi-autonomous learning method is designed to explain as much as possible of the dataset in one or multiple models within a predefined statistical space taking into account the ligand interaction space.

MODELS The semi-autonomous learning method identified three unique to partially overlapping robust LIE regression models.

3 automatically resolved regression models. 1 unique, 2 partly overlapping

86% of the dataset explained

01 Robust LIE regression models. Correlation plots (units in kJ mol⁻¹)

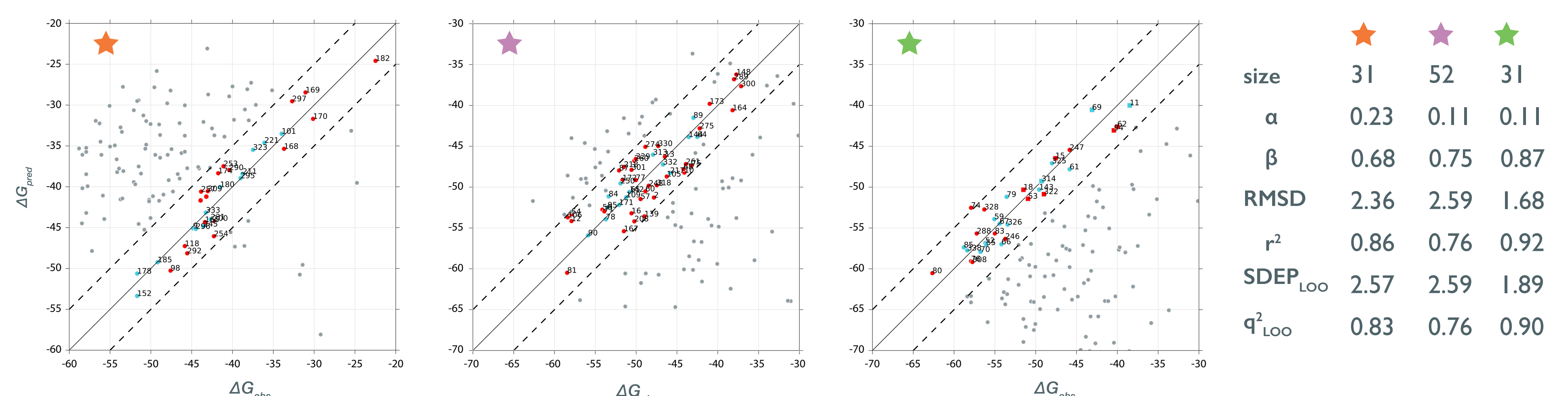|	★ (orange)	★ (purple)	★ (green)
size	31	52	31
α	0.23	0.11	0.11
β	0.68	0.75	0.87
RMSD	2.36	2.59	1.68
r²	0.86	0.76	0.92
SDEP_LOO	2.57	2.59	1.89
q²_LOO	0.83	0.76	0.90

INTERACTIONS Interaction profiles provided the applicability domain of the three models in terms of the type of interactions the ligands have with the four interaction "hotspots" ●

1 ★ model with steroid based aromatase inhibitors (SAI)

2 ★★ models with non-steroid based aromatase inhibitors (NSAI)

02 Interaction profiling. Contact frequency for different interaction types as horizontally stacked bars

Hydrophobic contacts
■ general   ■ π-stacking   □ T-stacking
■ π-cation   ■ halogen bond   ■ potential heme coordination

Salt bridge
■ neg. ligand - pos. protein   ■ pos. ligand - neg. protein

Hydrogen bond, ligand donor - protein acceptor
■ ligand aromatic N-atom   ■ ligand N-atom   ■ ligand O-atom
Hydrogen bond, ligand acceptor - protein donor
■ ligand aromatic N-atom   ■ ligand N-atom   ■ ligand O-atom

CONCLUSIONS Our LIE workflow provides binding affinity prediction with detailed structural information while keeping CPU costs tractable. Extensive statistical modelling and interaction profiling provides models that explain a maximum of the dataset with accompanying applicability domain and distribution confidence information.

Q225, D309, L477, S478, T310, ASD, M374, HEME, R115

[1] Korb, O. et al (2007). Swarm Intelligence, 1(2), pp. 115-134
[2] Malde AK, et al (2011). J. Chem. Theo. Comput., 7(12), 4026-4037
[3] Stjernschantz, E. et al (2006). J. Chem. Inf. Model, 46(5):1972-1983
[4] Vosmeer C.R. et al (2014). Int. J. Mol. Sci., 15:798-816
[5] Capoferri, L. et al (2015). PLoS ONE, 10(11)
[6] Pronk, S. et al (2013). Bioinformatics, 29(7):845-854