Universiteit Utrecht

HADDOCK High-Ambiguity Driven Docking

Bridging artificial intelligence and physics-based docking for better modelling of biomolecular complexes

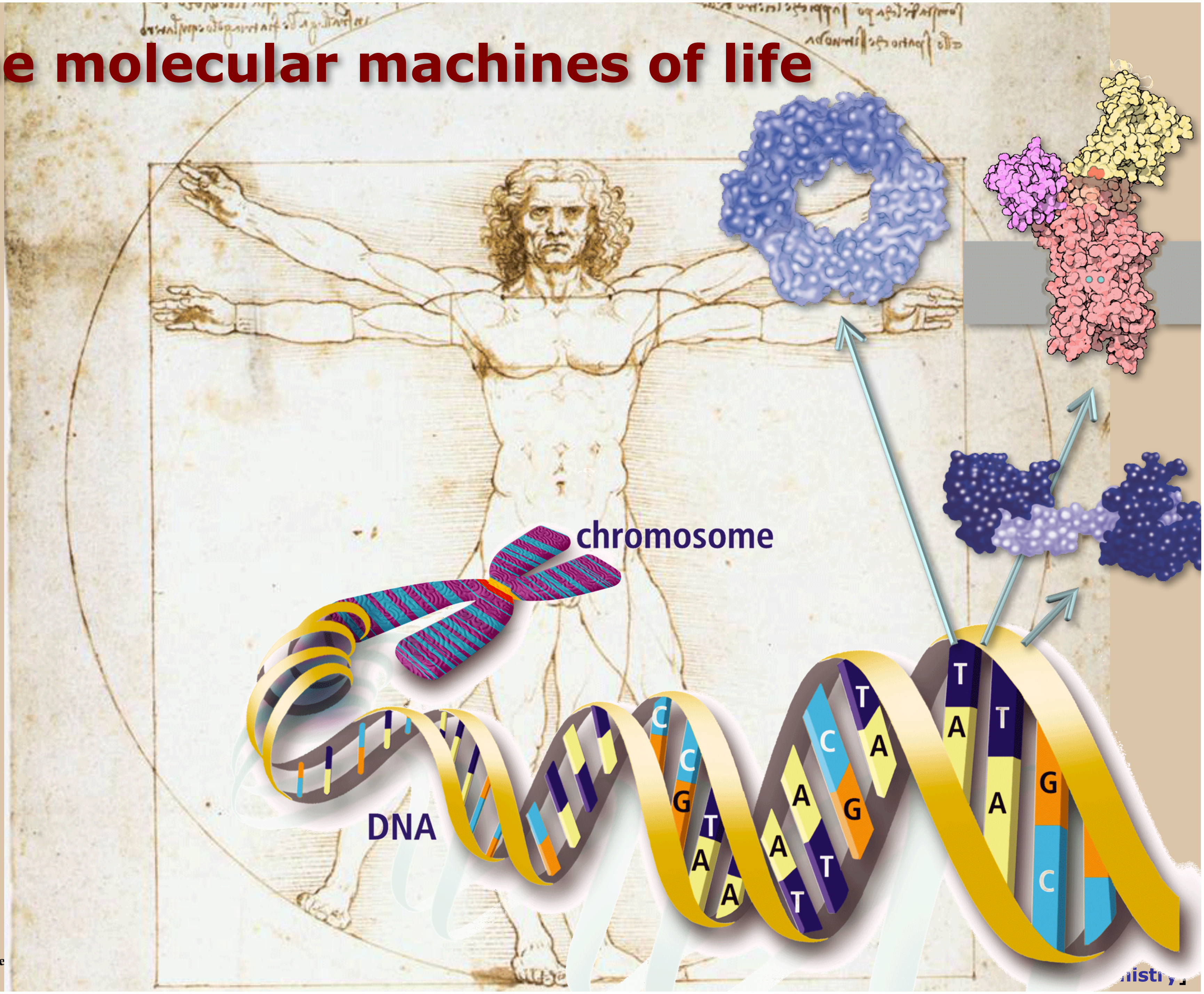
Li Xue

Computational Structural Biology Lab

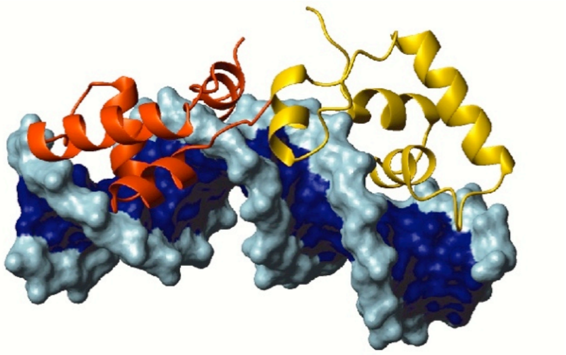
Utrecht University, the Netherlands

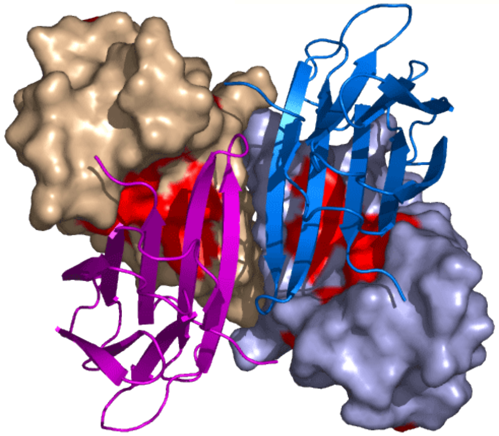
ISGC I Mar. 20th 2018 I Academia Sinica, Taipei

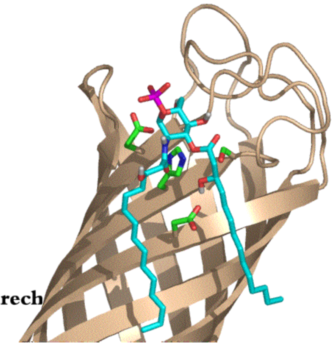
The molecular machines of life

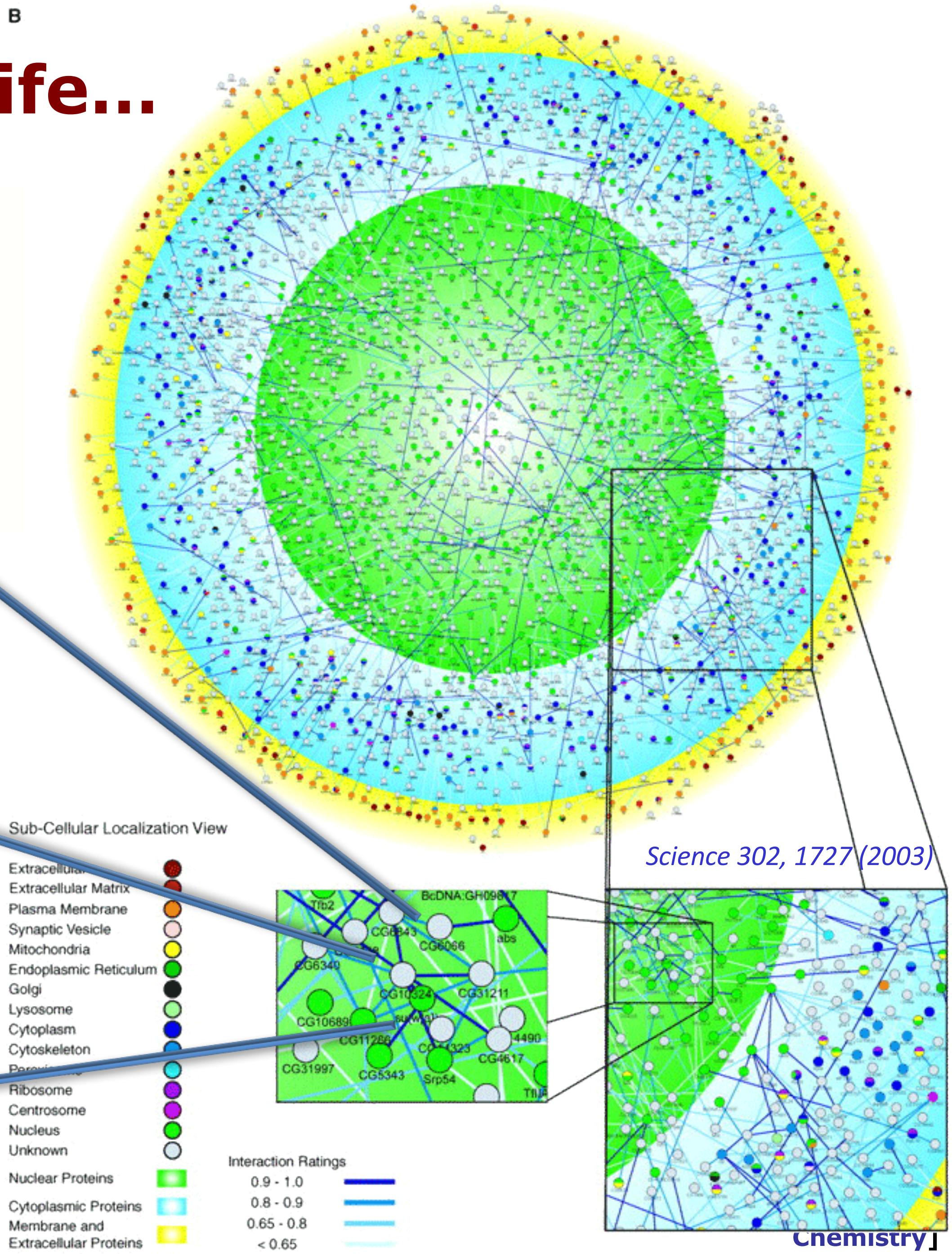


The network of life…









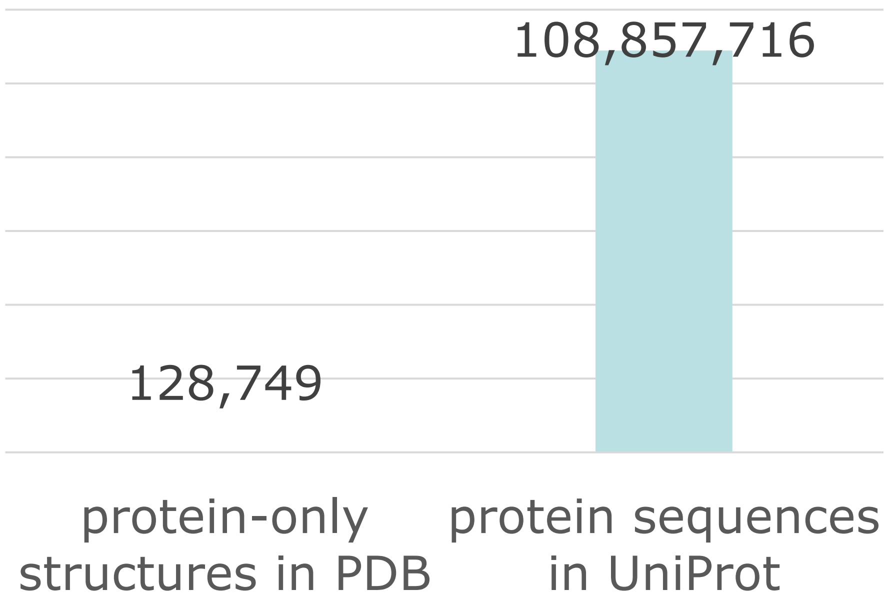
Science 302, 1727 (2003)

The Quest for 3D Structures of Protein Complexes

Structures/Models

- **Provide structural insight into protein function and regulation.**
- **Can guide experimental studies**
- **Can help rationalize the effect of genetic defects**
- **Can function as drug target**

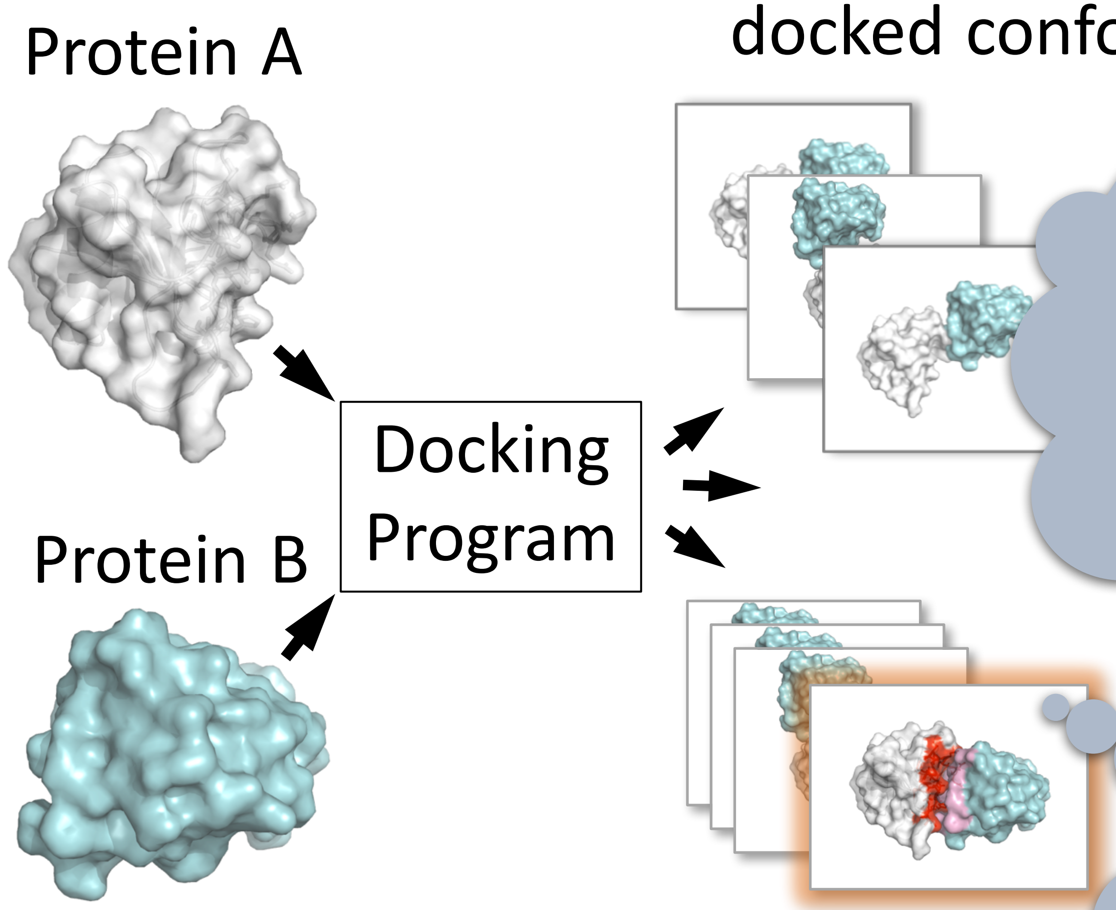
BUT: The number of experimentally solved protein structures are greatly lagged behind of sequences



This calls for complementary computational methods

Protein-Protein Docking & its challenges

Docking generates *vast numbers* of docked conformations



Protein A

Docking Program

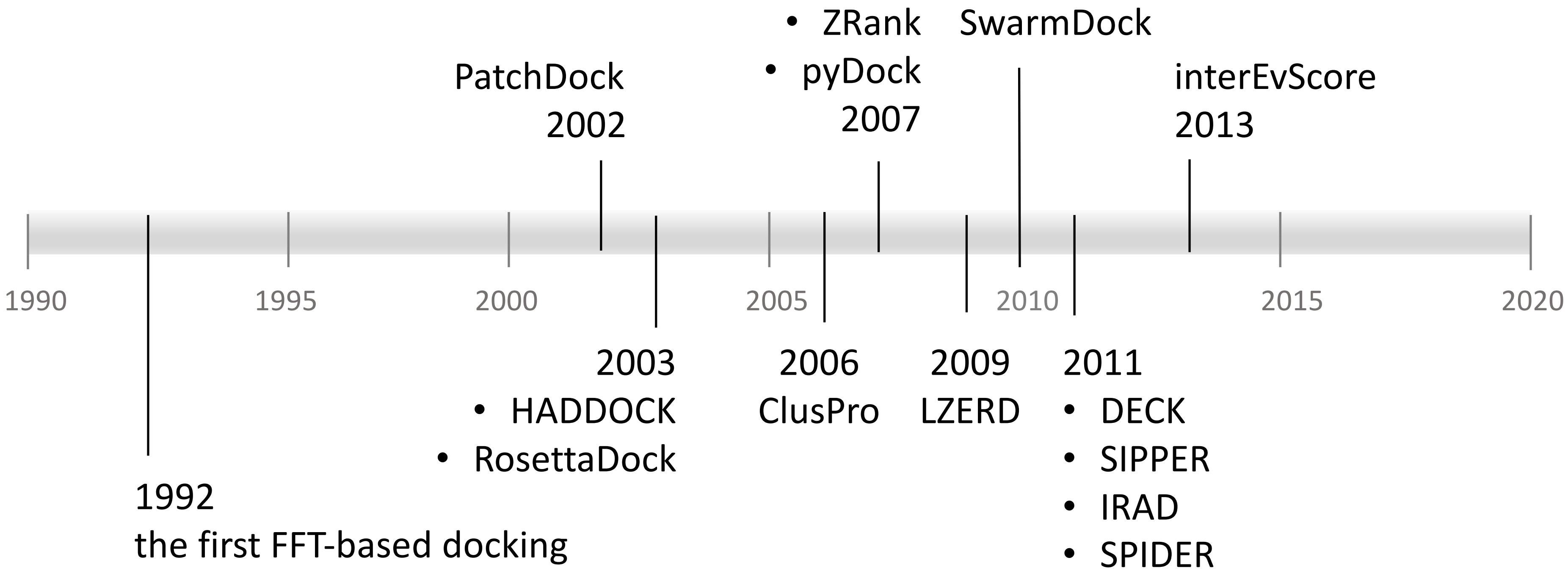
Protein B

Challenge 1 – Huge sampling space: How to generate native-like (correct) models for *flexible* proteins?

Challenge 2 – Scoring: How to *select* the native-like models?

Scoring in the past 25 years

> 100 scoring functions published



1992
the first FFT-based docking

PatchDock
2002

2003
- HADDOCK
- RosettaDock

2006
ClusPro

- ZRank
- pyDock

2007

2009
LZERD

SwarmDock

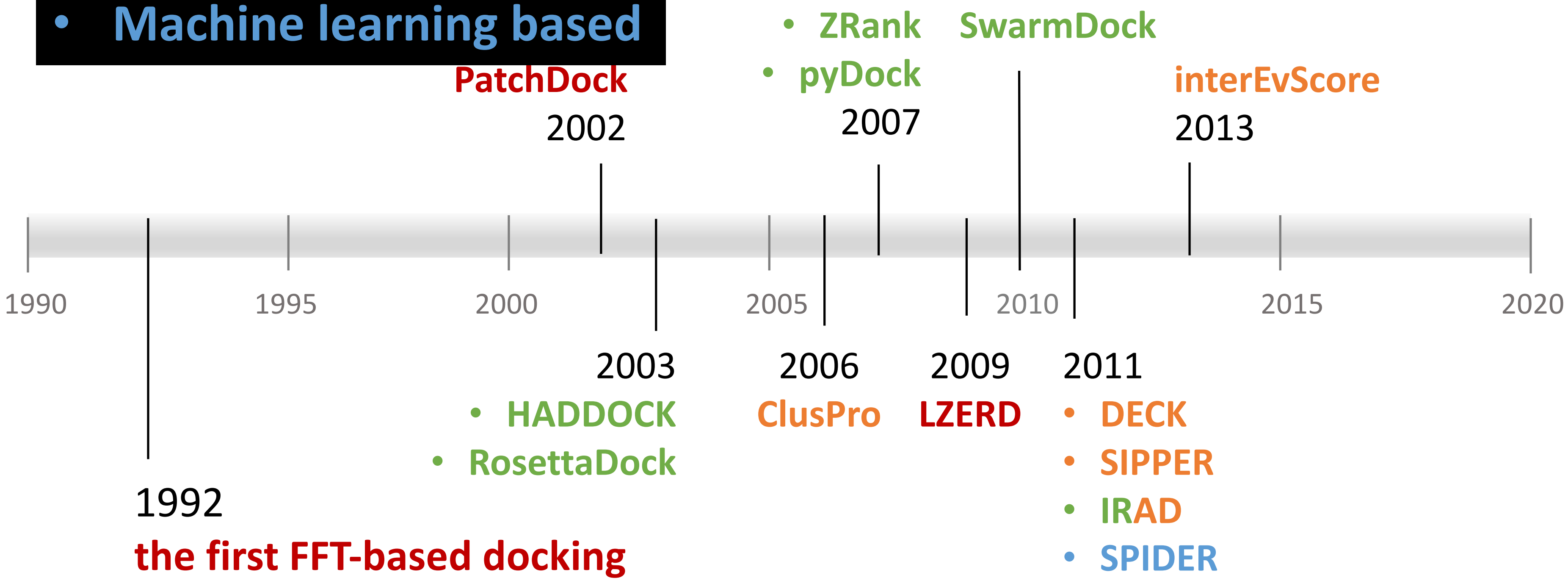
2011
- DECK
- SIPPER
- IRAD
- SPIDER

interEvScore
2013

Scoring in the past 25 years

> 100 scoring functions published

- Shape complementarity
- physics based
- Statistical potentials
- Machine learning based



1992
the first FFT-based docking

PatchDock
2002

2003
- HADDOCK
- RosettaDock

2006
ClusPro

- ZRank
- pyDock

2007

2009
LZERD

SwarmDock

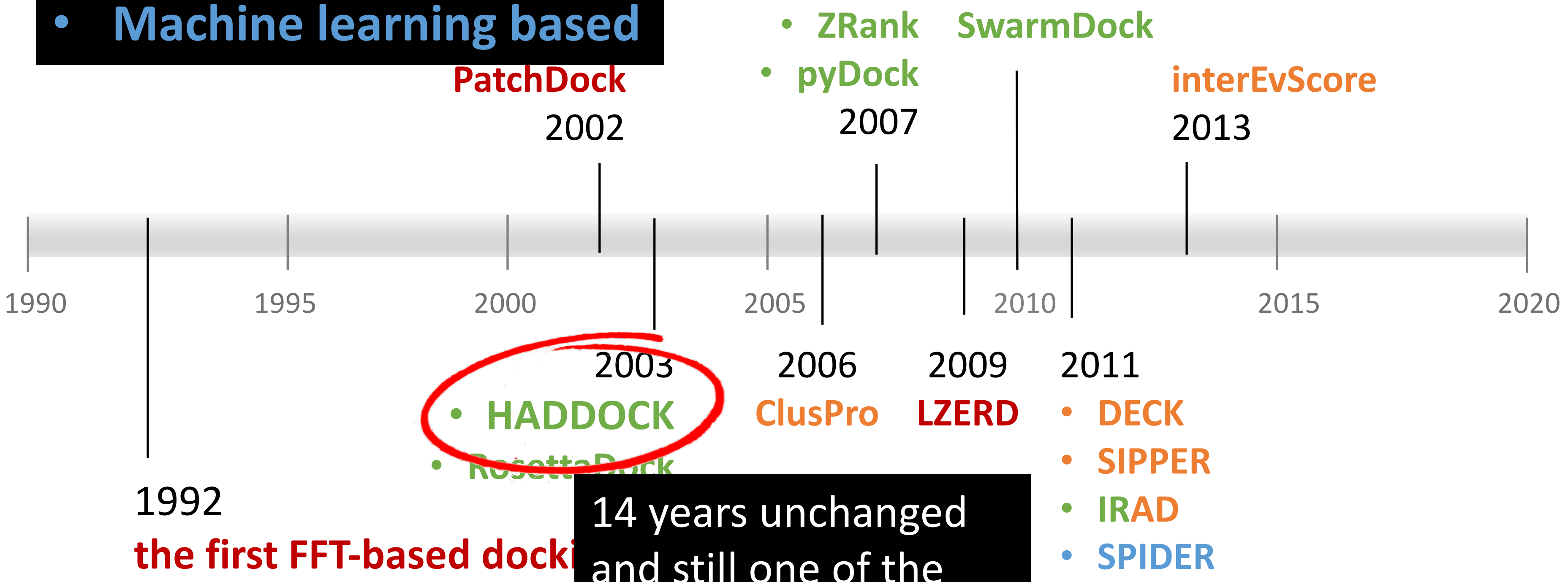
2011
- DECK
- SIPPER
- IRAD
- SPIDER

interEvScore
2013

Scoring in the past 25 years

> 100 scoring functions published

- Shape complementarity
- physics based
- Statistical potentials
- Machine learning based



1992
the first FFT-based docki

PatchDock
2002

2003
- HADDOCK
- RosettaDock

2006
ClusPro

- ZRank
- pyDock
2007

2009
LZERD

SwarmDock

2011
- DECK
- SIPPER
- IRAD
- SPIDER

interEvScore
2013

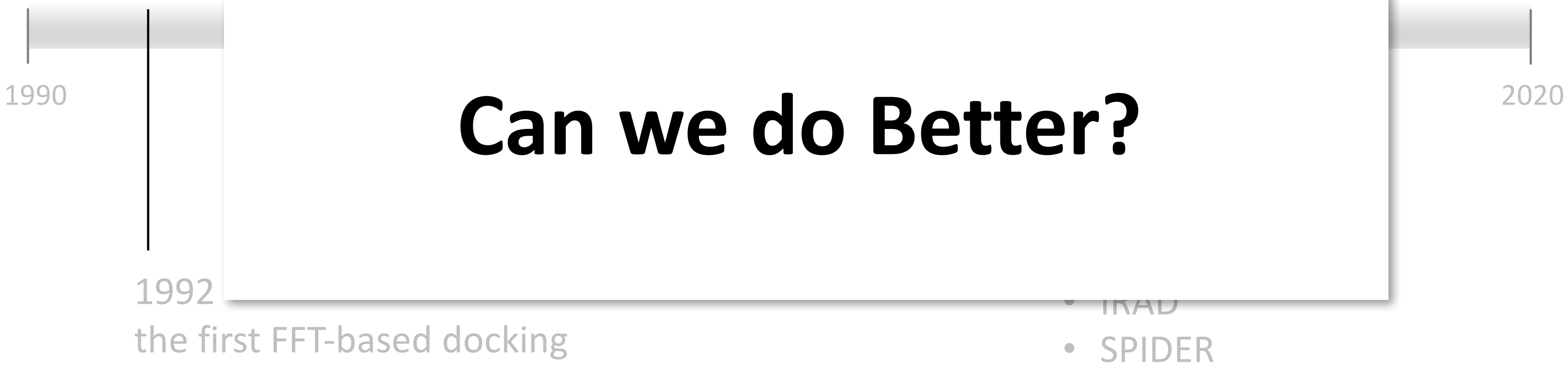
14 years unchanged and still one of the best.

Scoring in the past 25 years

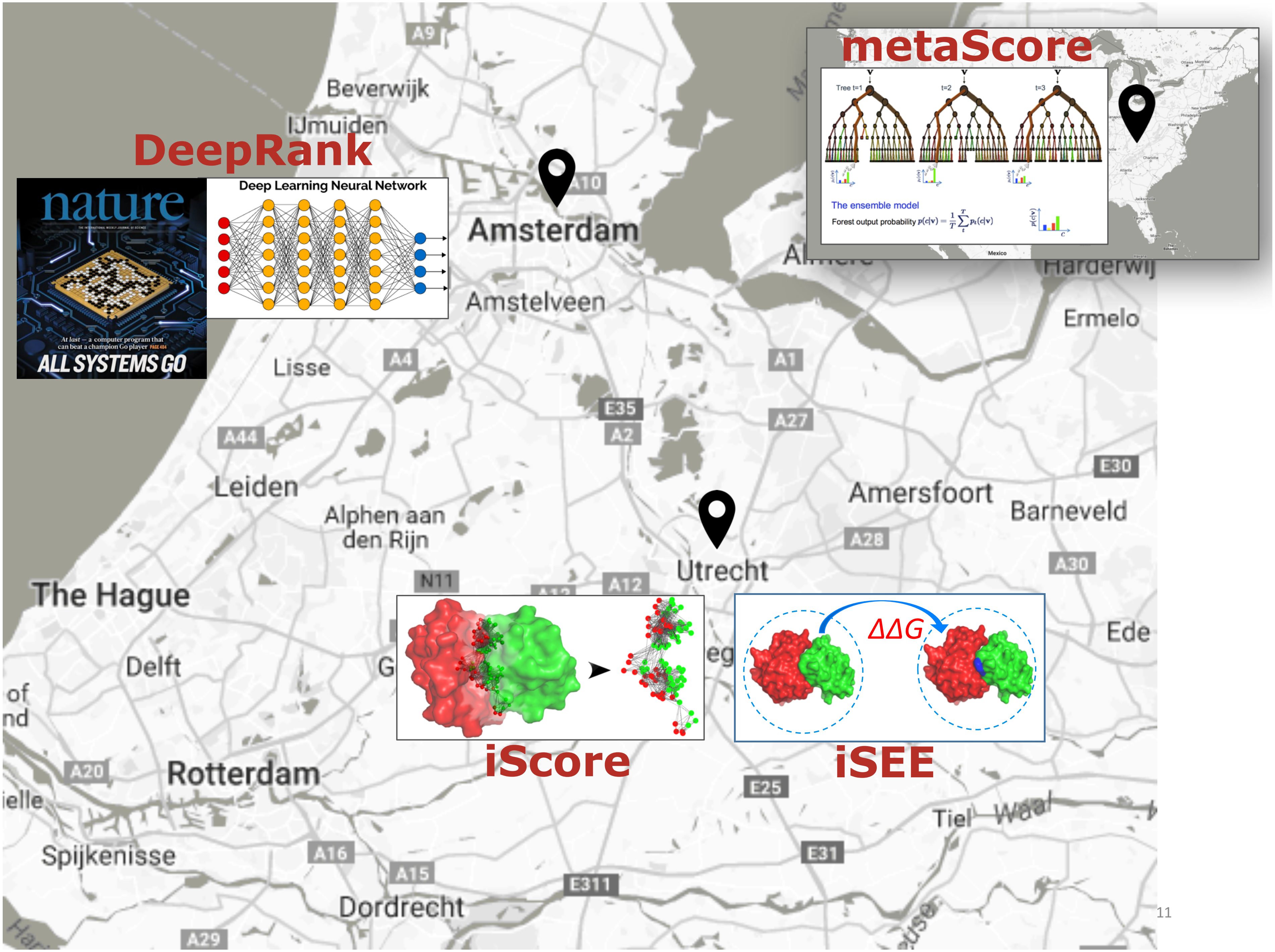
> 100 scoring functions published

- More data accumulated
- More computational power
- Better machine learning algorithms

Can we do Better?



netherlands
eScience center
PENNSTATE
1 8 5 5
United States
Mexico
Beverwijk
IJmuiden
Amsterdam
Amstelveen
Almere
Harderwijk
Ermelo
Lisse
Leiden
Alphen aan den Rijn
Amersfoort
Barneveld
Utrecht
Ede
The Hague
Delft
Gouda
Rotterdam
Spijkenisse
Dordrecht
A9
A10
A5
A4
A1
A44
E35
A2
A27
E30
A28
A30
N11
A12
A20
A16
A15
E311
E31
A29
mutagenesis
cryo-EM
nmr
xl-ms
saxs
HADDOCK
High-Ambiguity Driven Docking



DeepRank

metaScore

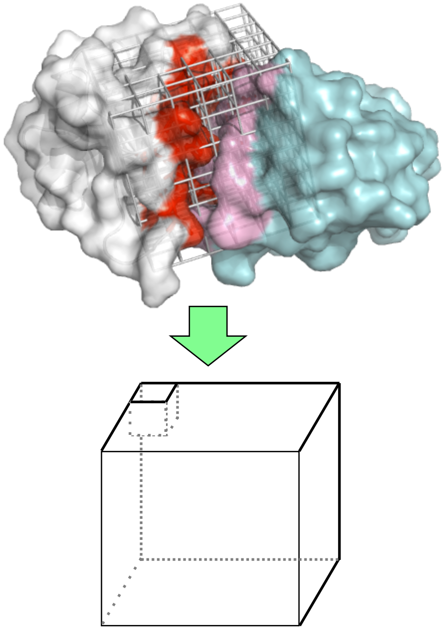
iScore

iSEE

Our DeepRank project:
Why Deep Learning, specifically convNets?

Human expert visually checking the interface of a docked model
@ CAPRI 38 competition – manual selection stage

A docked model
with its interface (red/pink) in a grid box



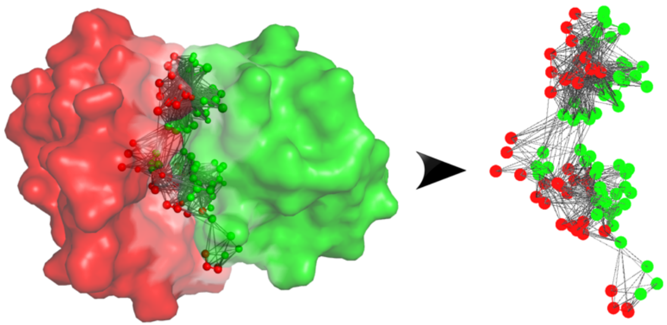
Use convNets to scan the interface of a docked model

https://github.com/DeepRank

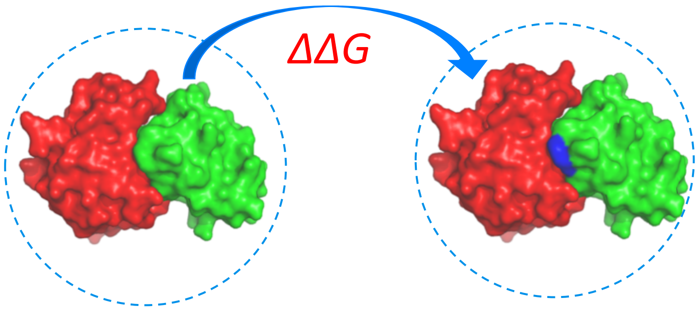
STAYTUNED

Outline

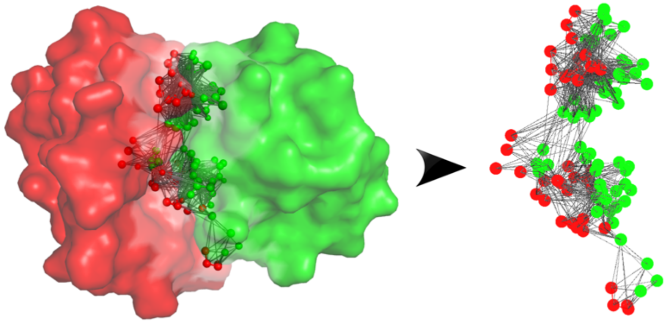
iScore



iSEE



structure → interactions → binding free energy → function



iScore

interface graph based docking scoring function

structure → interactions → binding free energy → function

iScore: a novel machine learning based scoring function

I propose a novel approach that treat the scoring problem as a *graph comparison problem.*

A docked model and its interface graph

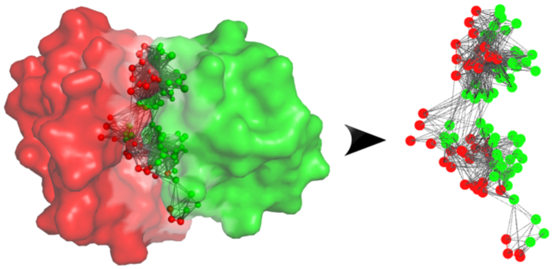


Figure credit: Cunliang Geng

Calculate graph similarity

Training data

Positive data

Interface graphs from *correct* protein-protein structures

Negative data

Interface graphs from *wrong* protein-protein models

iScore: a novel machine learning based scoring function

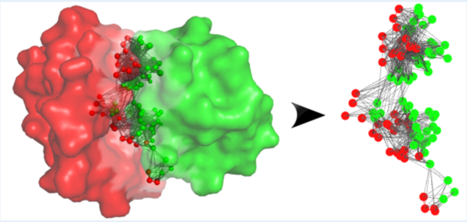
***Strength #1*: Much higher resolution of the input information is considered.**

Typical machine learning scoring function :

Uses ***whole interface descriptors*** *instead of* working at the resolution of atoms or residues.

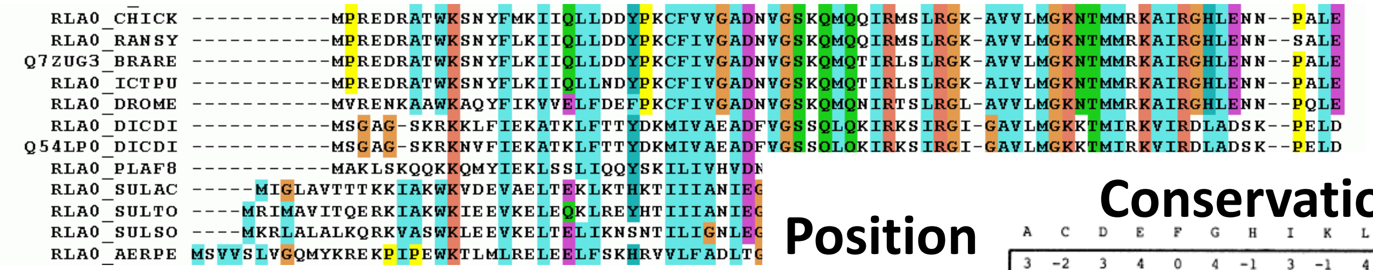
iScore exploits:

- Network topology
- Atom/Residue level information (node label)
- Pairwise information (edge label)



iScore: a novel machine learning based scoring function

Strength #2: Full profile of interface conservation can be exploited.



```
    RLA0_CHICK -----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
    RLA0_RANSY -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE
  Q7ZUG3_BRARE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
    RLA0_ICTPU -----------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE
    RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE
    RLA0_DICDI -----------MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD
  Q54LP0_DICDI -----------MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD
    RLA0_PLAF8 -----------MAKLSKQQKKQMYIEKLSSLIQQYSKILIVHVDN
    RLA0_SULAC -----MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEG
    RLA0_SULTO ----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEG
    RLA0_SULSO ----MKRLALALKQRKVASWKLEEVKELTELIKNSNTILIGNLEG
    RLA0_AERPE MSVVSLVGQMYKREKPIPEWKTLMLRELEELFSKHRVVLFADLTG
```

Position

1

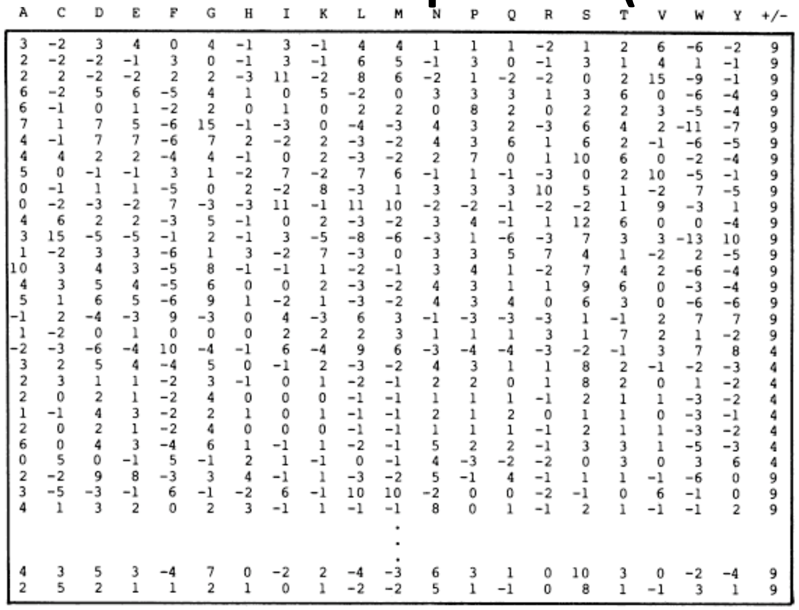
2

3

4

…

Conservation profiles (a PSSM)



A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	+/-
3	-2	3	4	0	4	-1	3	-1	4	4	1	1	1	-2	1	2	6	-6	-2	9
2	-2	-2	-1	3	0	-1	3	-1	6	5	-1	3	0	-1	3	1	4	1	-1	9
2	2	-2	-2	2	2	-3	11	-2	8	6	-2	1	-2	-2	0	2	15	-9	-1	9
6	-2	5	6	-5	4	1	0	5	-2	0	3	3	3	1	3	6	0	-6	-4	9
6	-1	0	1	-2	2	0	1	0	2	2	0	8	2	0	2	2	3	-5	-4	9
7	1	7	5	-6	15	-1	-3	0	-4	-3	4	3	2	-3	6	4	2	-11	-7	9
4	-1	7	7	-6	7	2	-2	2	-3	-2	4	3	6	1	6	2	-1	-6	-5	9
4	4	2	2	-4	4	-1	0	2	-3	-2	2	7	0	1	10	6	0	-2	-4	9
5	0	-1	-1	3	1	-2	7	-2	7	6	-1	1	-1	-3	0	2	10	-5	-1	9
0	-1	1	1	-5	0	2	-2	8	-3	1	3	3	3	10	5	1	-2	7	-5	9
0	-2	-3	-2	7	-3	-3	11	-1	11	10	-2	-2	-1	-2	-2	1	9	-3	1	9
4	6	2	2	-3	5	-1	0	2	-3	-2	3	4	-1	1	12	6	0	0	-4	9
3	15	-5	-5	-1	2	-1	3	-5	-8	-6	-3	1	-6	-3	7	3	3	-13	10	9
1	-2	3	3	-6	1	3	-2	7	-3	0	3	3	5	7	4	1	-2	2	-5	9
10	3	4	3	-5	8	-1	-1	1	-2	-1	3	4	1	-2	7	4	2	-6	-4	9
4	3	5	4	-5	6	0	0	2	-3	-2	4	3	1	1	9	6	0	-3	-4	9
5	1	6	5	-6	9	1	-2	1	-3	-2	4	3	4	0	6	3	0	-6	-6	9
-1	2	-4	-3	9	-3	0	4	-3	6	3	-1	-3	-3	-3	1	-1	2	7	7	9
1	-2	0	1	0	0	0	2	2	2	3	1	1	1	3	1	7	2	1	-2	9
-2	-3	-6	-4	10	-4	-1	6	-4	9	6	-3	-4	-4	-3	-2	-1	3	7	8	4
3	2	5	4	-4	5	0	-1	2	-3	-2	4	3	1	1	8	2	-1	-2	-3	4
2	3	1	1	-2	3	-1	0	1	-2	-1	2	2	0	1	8	2	0	1	-2	4
2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
1	-1	4	3	-2	2	1	0	1	-1	-1	2	1	2	0	1	1	0	-3	-1	4
2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
6	0	4	3	-4	6	1	-1	1	-2	-1	5	2	2	-1	3	3	1	-5	-3	4
0	5	0	-1	5	-1	2	1	-1	0	-1	4	-3	-2	-2	0	3	0	3	6	4
2	-2	9	8	-3	3	4	-1	1	-3	-2	5	-1	4	-1	1	1	-1	-6	0	9
3	-5	-3	-1	6	-1	-2	6	-1	10	10	-2	0	0	-2	-1	0	6	-1	0	9
4	1	3	2	0	2	3	-1	1	-1	-1	8	0	1	-1	2	1	-1	-1	2	9
										⋮										
4	3	5	3	-4	7	0	-2	2	-4	-3	6	3	1	0	10	3	0	-2	-4	9
2	5	2	1	1	2	1	0	1	-2	-2	5	1	-1	0	8	1	-1	3	1	9

Evolution information is critical for

- **Protein recognition**
- **Protein folding**

iScore v1.0: an evolution based scoring function

iScore v1.0:

- uses residue conservation only.
- Residue level resolution

Trained on 114 bound complexes, 114 wrong models.

Tested on 64 cases, 400 docked models per case.

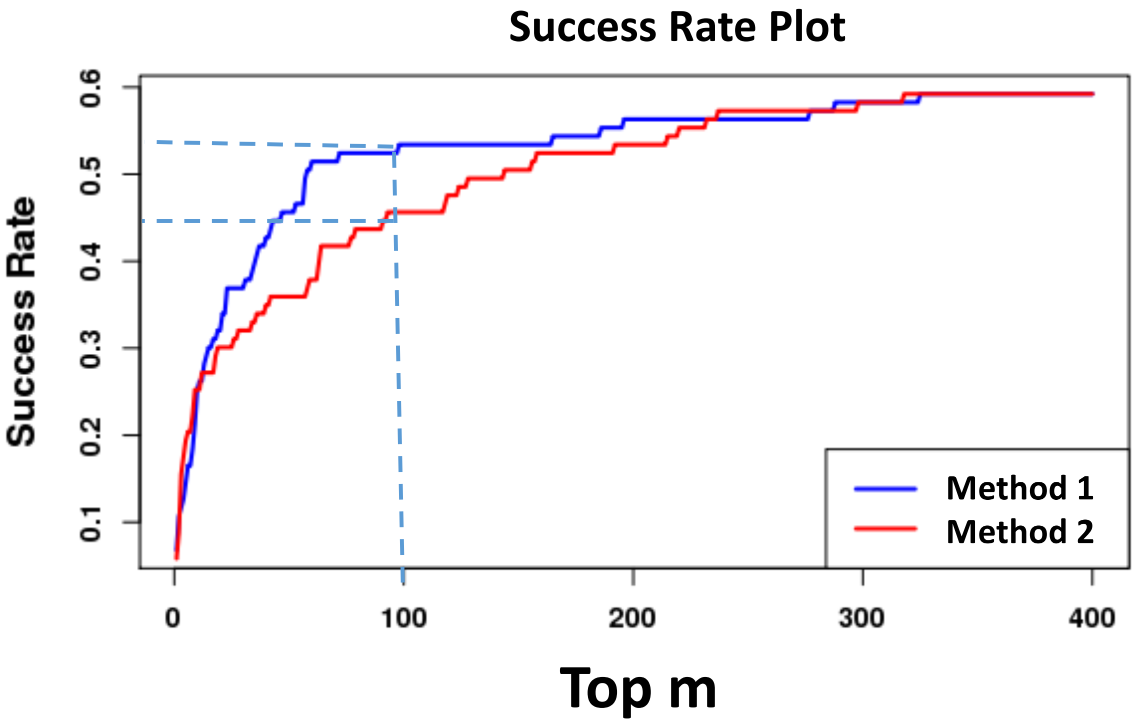
iScore vs. HADDOCK score

HADDOCK score	iScore v1.0
Atom resolution	Residue resolution
Interaction cutoff: 8.5 Å	Interaction cutoff: 6 Å
Linear function	Non-linear graph similarity based
Interaction energy based	Residue conservation based
Fast: 2 seconds per model	Slow: 0.05-0.5 second per graph pair , but many comparisons

Evaluation: success rate and hit rate

A hit: a correct (near-native) docked model (interface RMSD ≤ 4 Å)

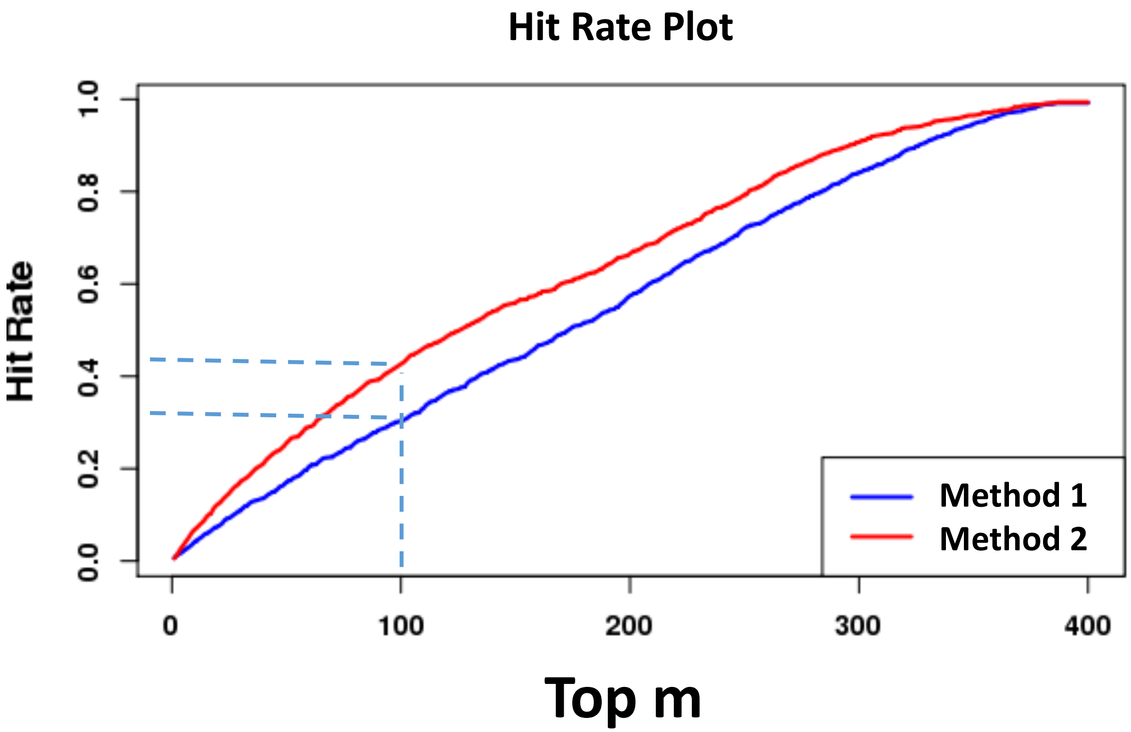
Success Rate:
the percentage of cases that have at least one hit among the top *m* conformations.



Evaluation: success rate and hit rate

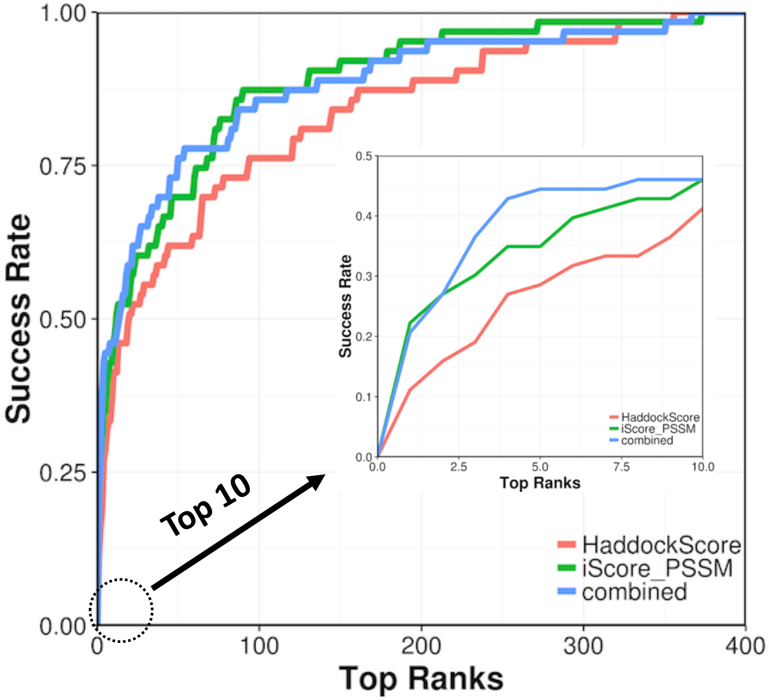
A hit: a correct (near-native) docked model (interface RMSD ≤ 4 Å)

Average Hit Rate:
the percentage of hits that are included among the top *m* conformations.

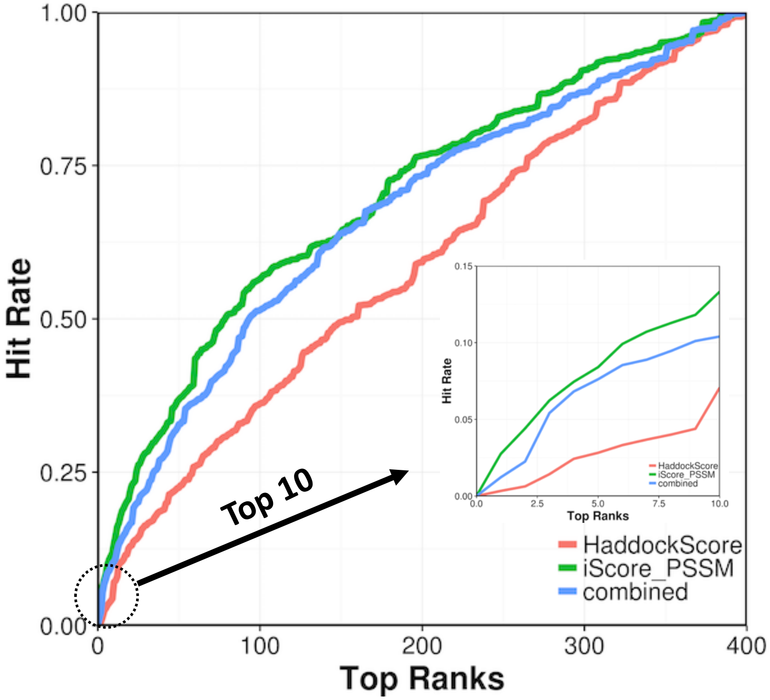


iScore vs. HADDOCK score

Success rate



Hit rate



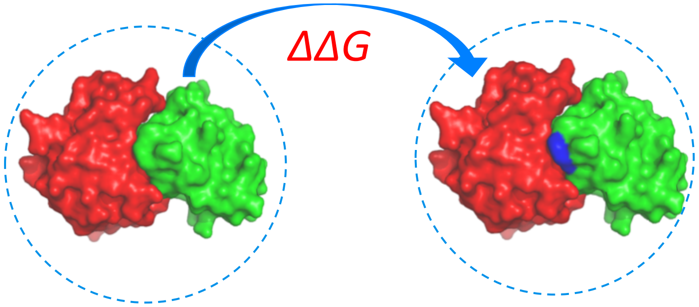
The Combined Score is the average of iScore and Haddock Score.

Summary of iScore

- iScore is an effective graph based scoring function.
- iScore v1.0 (*based on conservation only*) outperforms HADDOCK scoring function on the majority of dimer cases of docking benchmark 4.0.

iScore complements HADDOCK score:

HADDOCK score = 1.0 Evdw + 0.2 Eelec + 1.0 Edesolv



iSEE

interface **S**tructure, **E**nergy and **E**volution based ΔΔG predictor

structure → interactions → binding free energy → function

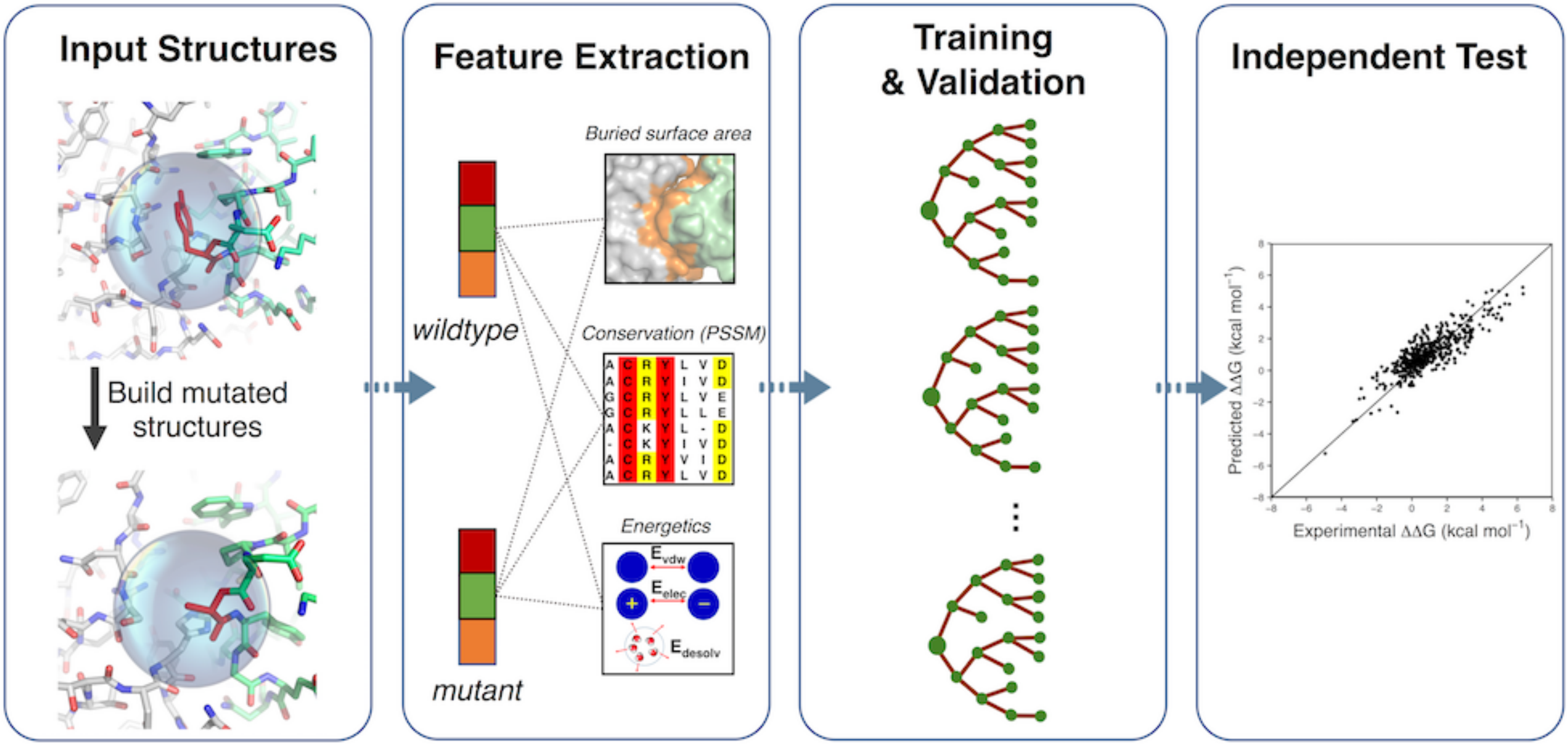
Why are we interested in mutations?

- Coding variants -> phenotype
- Carcinogens
- Errors in reading tRNAs at the ribosome
- The generation of antigen-binding CDR loops of antibodies
- Protein engineering
- Directed evolution

iSEE

iSEE aims to model:

Binding affinity change upon mutation (ΔΔG) = f (features of the mutation site)

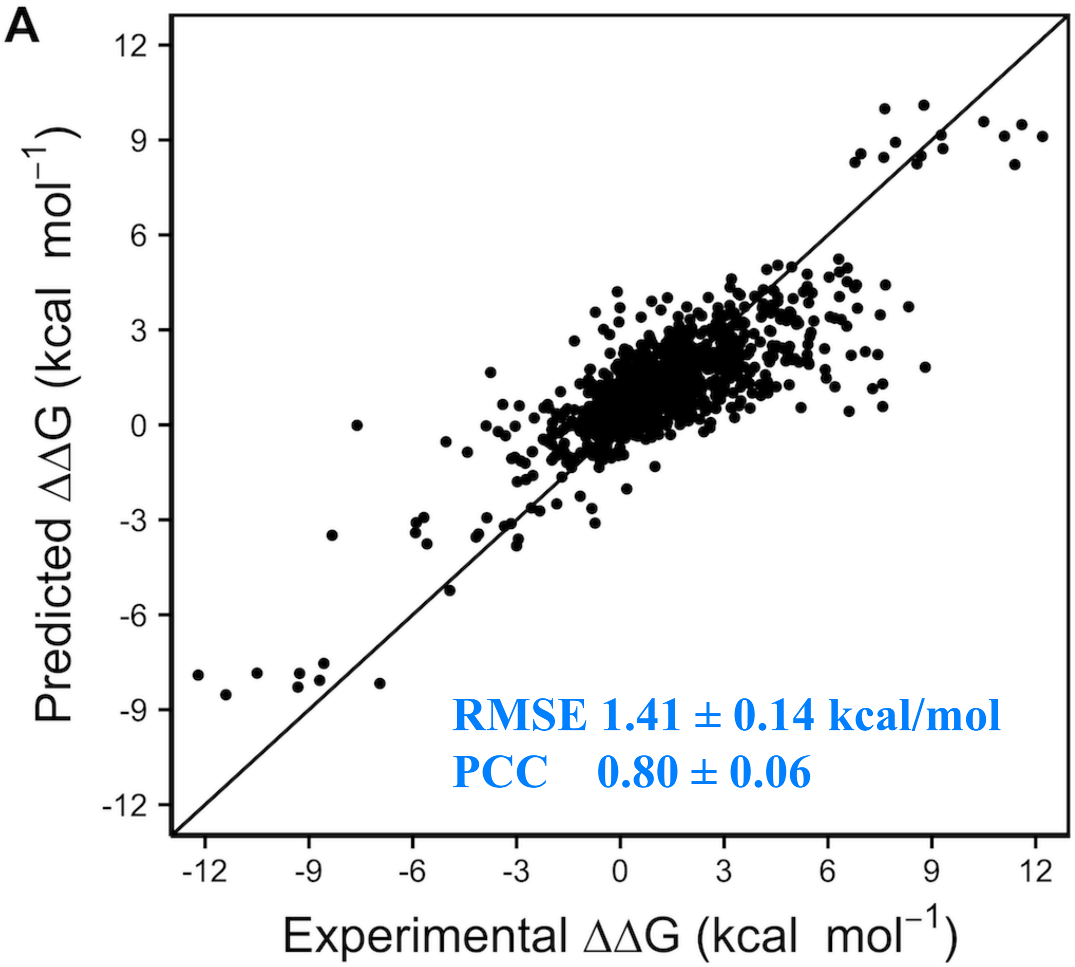


A	C	R	Y	L	V	D
A	C	R	Y	I	V	D
G	C	R	Y	L	V	E
G	C	R	Y	L	L	E
A	C	K	Y	L	-	D
-	C	K	Y	I	V	D
A	C	R	Y	V	I	D
A	C	R	Y	L	V	D

Geng et al. iSEE: Interface Structure, Evolution and Energy-based random forest predictor of binding affinity changes upon mutations, submitted to Bioinformatics.

10 times 10-fold cross validations

iSEE: **i**nterface **S**tructure, **E**nergy and **E**volution based ΔΔG predictor.



RMSE 1.41 ± 0.14 kcal/mol
PCC 0.80 ± 0.06

iSEE on two blind test datasets

The NM dataset (19 mutations)

Method	RMSE*	PCC
iSEE	**1.37**	**0.73**
BindProfX	**1.11**	**0.81**
FoldX	**1.15**	**0.72**
ZeMu	**1.28**	**0.70**
CC/PBSA[a]	1.33	0.60
pred2	1.44	0.48
pred1	1.49	0.39
BeAtMuSiC	1.70	0.24
mCSM	1.83	0.16

The MDM2-p53 dataset (33)

Method	RMSE*	PCC
iSEE	**0.77**	**0.65**
BindProfX	1.17	0.51
FoldX	1.36	0.50
BeAtMuSiC	1.02	0.48
mCSM	0.97	0.22

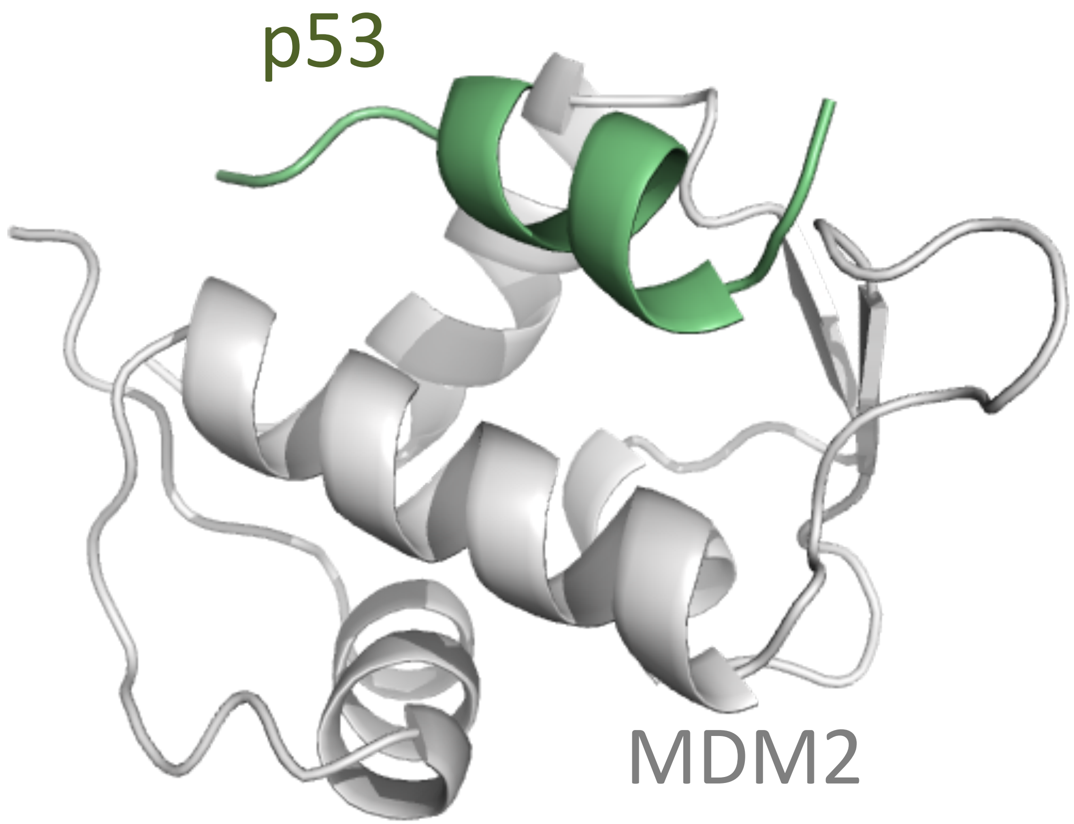
***RMSE in kcal mol^{-1}**

[a] The CC/PBSA predictor was trained on a dataset including the NM dataset.

iSEE competes favorably with state-of-the-art ΔΔG predictors on two blind test datasets.

Full mutation scanning on MDM2-p53 complex

MDM2-p53 is a prime target for cancer therapeutics

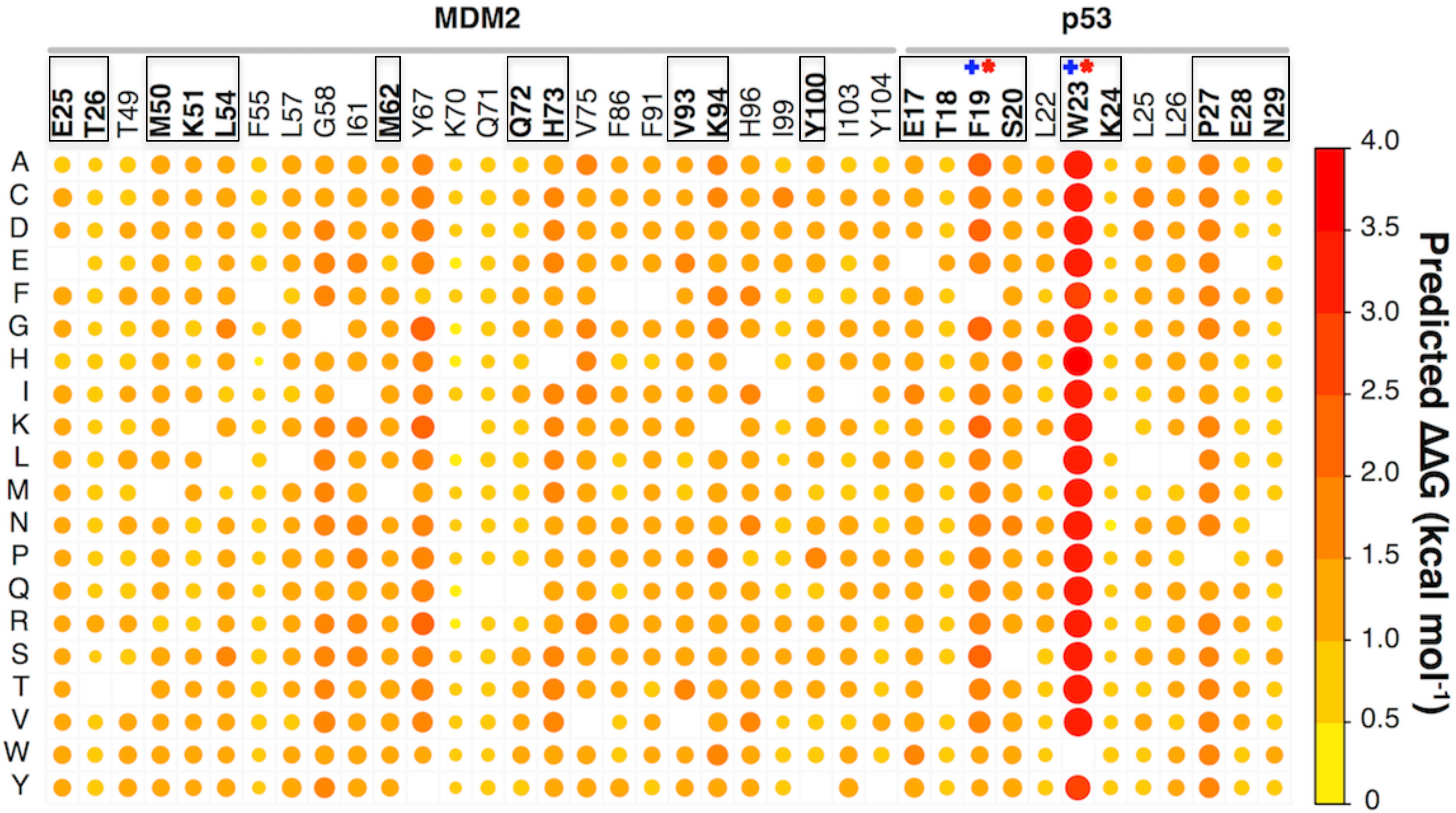


PDB ID: 1YCR

MDM2-p53 dataset

- **33 mutations: MDM2(16) and p53(17).**
- **With corresponding experimental $\Delta\Delta G$ data.**

Full mutation scanning on MDM2-p53 complex

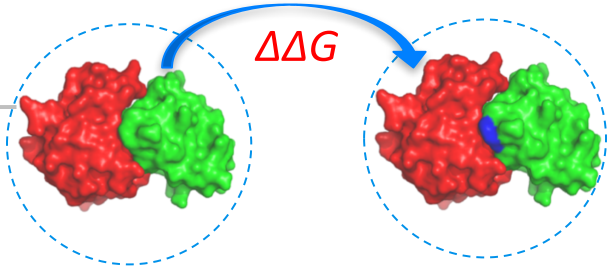


+ Experimental hot-spots

* Predicted hot-spots

iSEE enables high-throughput scanning.

Summary of iSEE

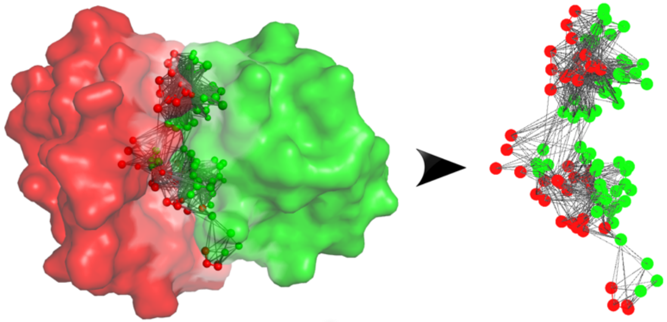


1. **iSEE is a machine learning-based ΔΔG predictor.**
2. **iSEE competes with state-of-the-art predictors.**
3. **iSEE can be used for high-throughput applications.**

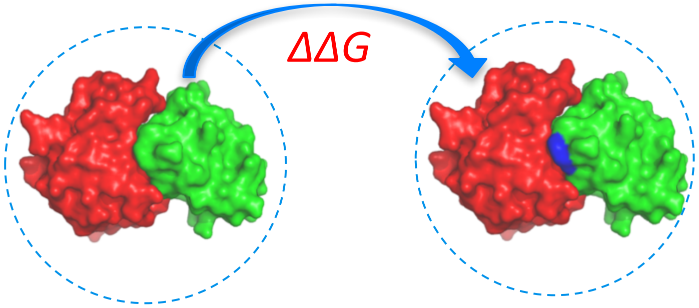
https://github.com/haddocking/iSee

Summary

iScore



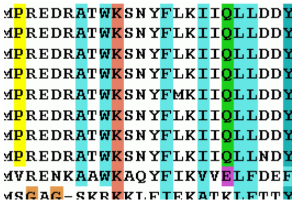
iSEE



structure → interactions → binding free energy → function

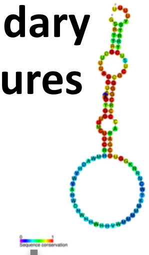
Artificial Intelligence Toward protein Interaction Prediction

Evolution information

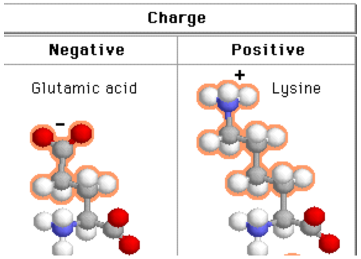


```
MPREDRATWKSNYFLKIIQLLDDY
MPREDRATWKSNYFLKIIQLLDDY
MPREDRATWKSNYFLKIIQLLDDY
MPREDRATWKSNYFLKIIQLLDDY
MPREDRATWKSNYFMKIIQLLDDY
MPREDRATWKSNYFLKIIQLLDDY
MPREDRATWKSNYFLKIIQLLDDY
MPREDRATWKSNYFLKIIQLLNDY
MVRENKAAWKAQYFIKVVELFDEF
MSGAG-SKRKKLFIEKATKLFTTY
```

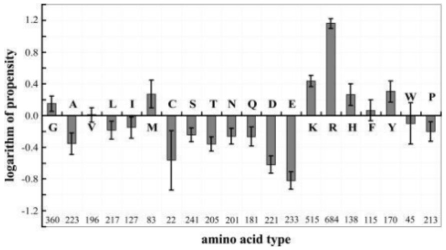
Secondary structures

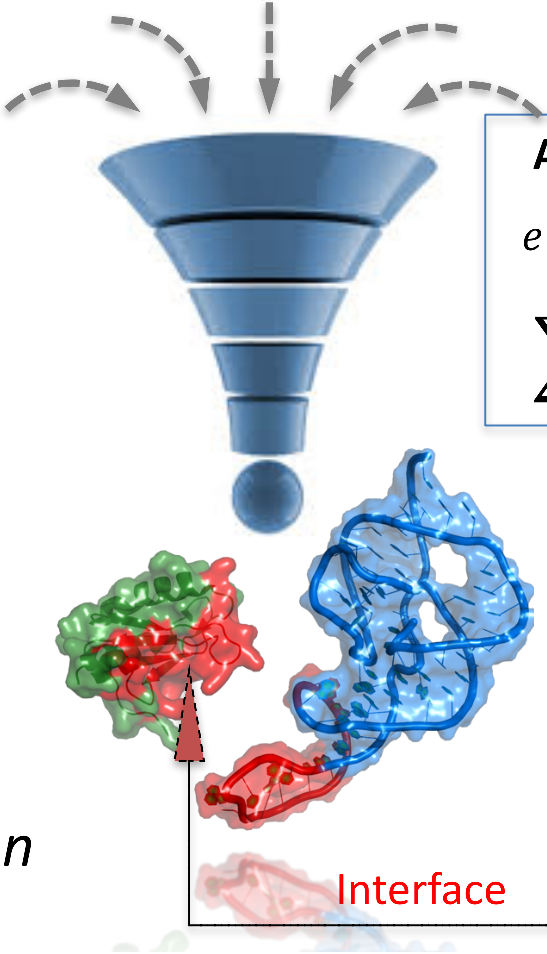


Physico-chemical attributes



Interface propensity





A novel interface potential score

$$e(w_P, w_R) = \min_{r \leq 5Å} e(A_n, R_n, r) +$$

$$\sum\nolimits_{i \neq n} \min_{j, r \leq 10Å} e(A_i, R_j, r)$$

Improved 3D modelling of protein interaction

Acknowledgement

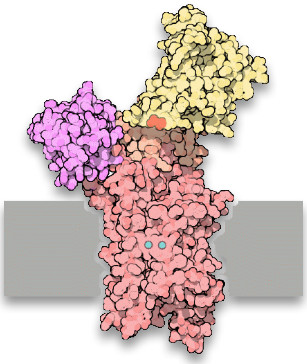
NWO
Netherlands Organisation for Scientific Research

netherlands
eScience center

West-Life

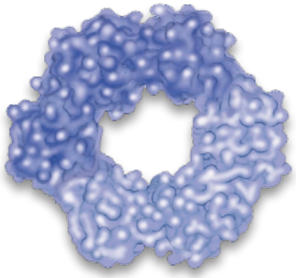
Thank You !

What can we learn from 3D structures (models) of complexes?

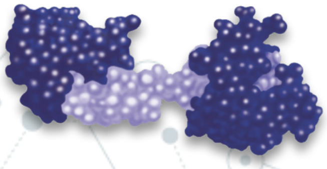


- **Models provide structural insight into function and mechanism of action**

- **Models can drive and guide experimental studies**



- **Models can help understand and rationalize the effect of disease-related mutations**



- **Models provide a starting point for drug design**

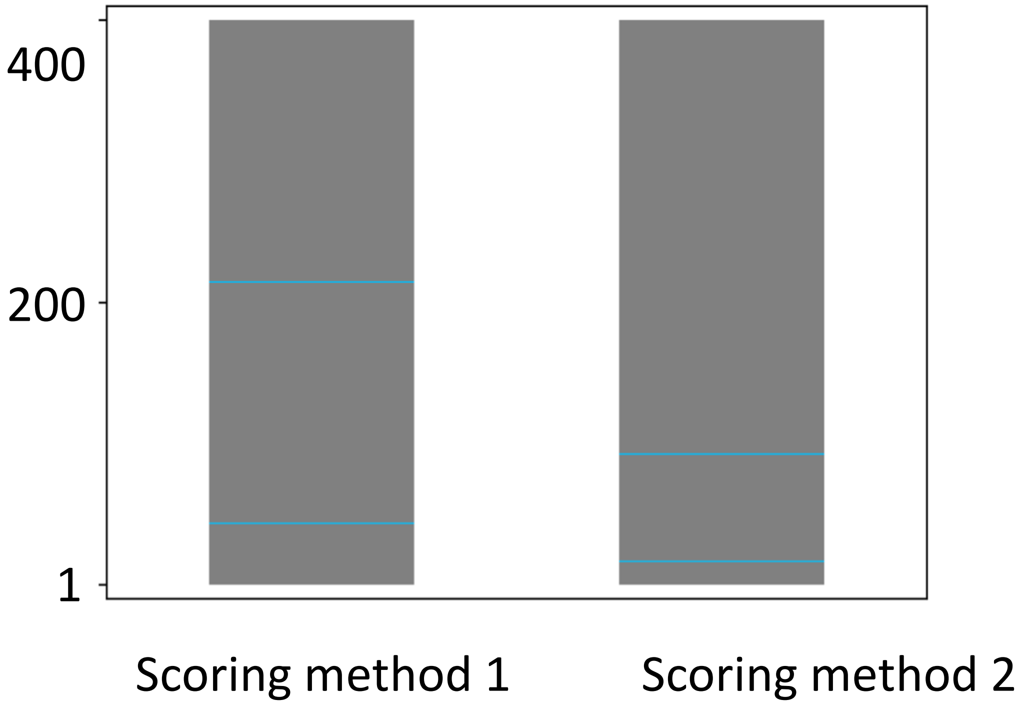
Why a reliable scoring function is so important?

- Docking
- Binding affinity estimation
- Hot-spot identification (ddG)
- Interaction design
- Interactome prediction

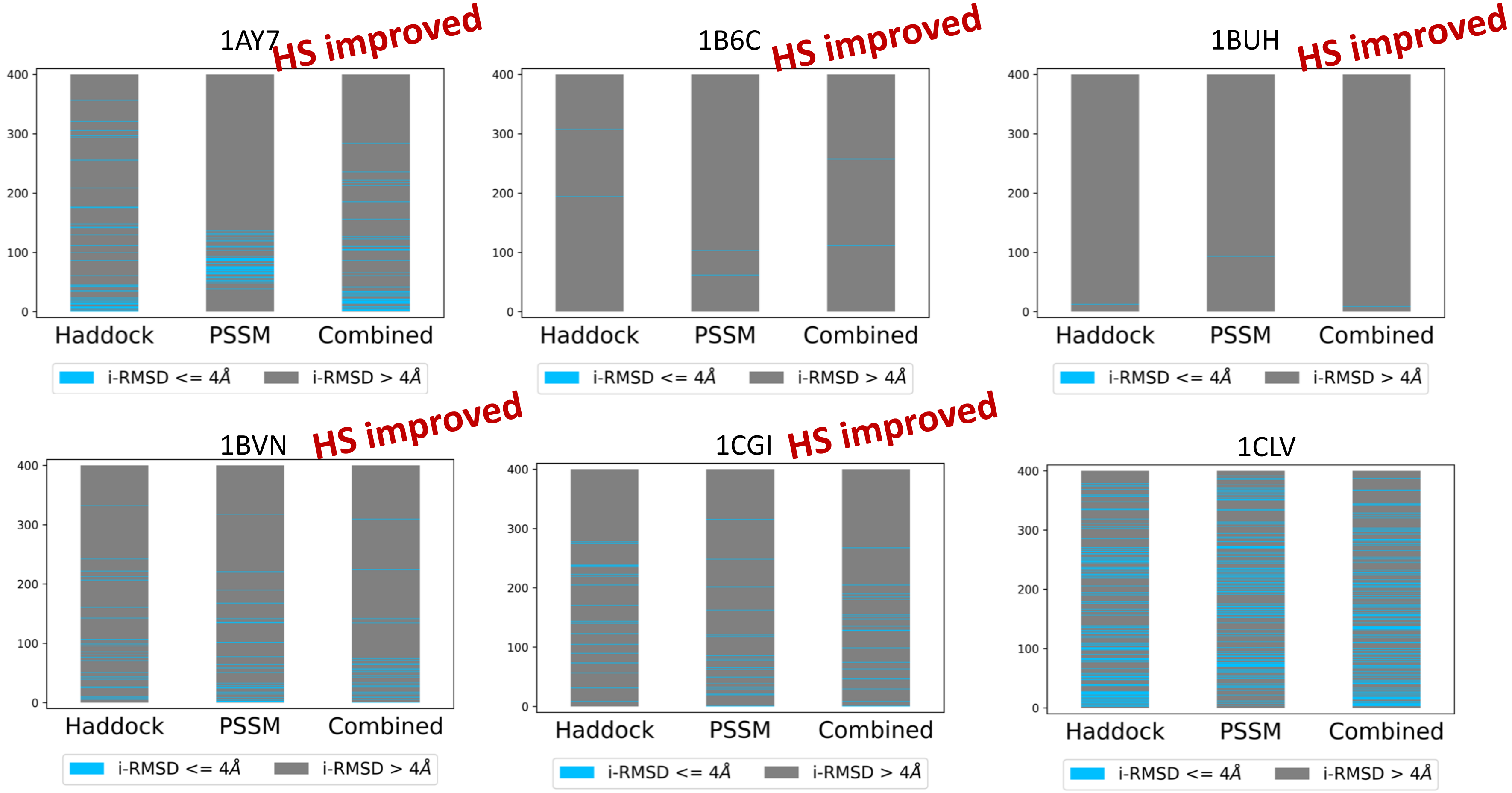
structure → interactions → binding free energy → function

Evaluation - melquiplot

y-axis:
Decoy Ranking
(the smaller the better)



iScore vs. HADDOCK score



Computational methods for predicting $\Delta\Delta G$

Rigorous methods, e.g., FEP

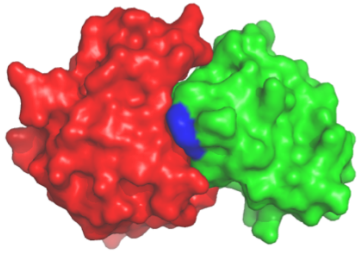
Accurate

However,

Requires extensive conformation sampling

Empirical energy based methods

Fast but not accurate



Machine learning based

Fast and can integrate heterogonous data

Current leading ΔΔG predictors

BindProfX **(Y Zhang et al., 2017)**

Interface structural profile + FoldX physics potentials

mCSM **(D Pires et al., 2014)**

contacts + pharmacophore change on mutation position

ZeMu **(D Dourado et al., 2014)**

MD simulations for flexibility of mutation zone + FoldX potentials

BeAtMuSiC **(Y Dehouck et al., 2013)**

Statistical potentials

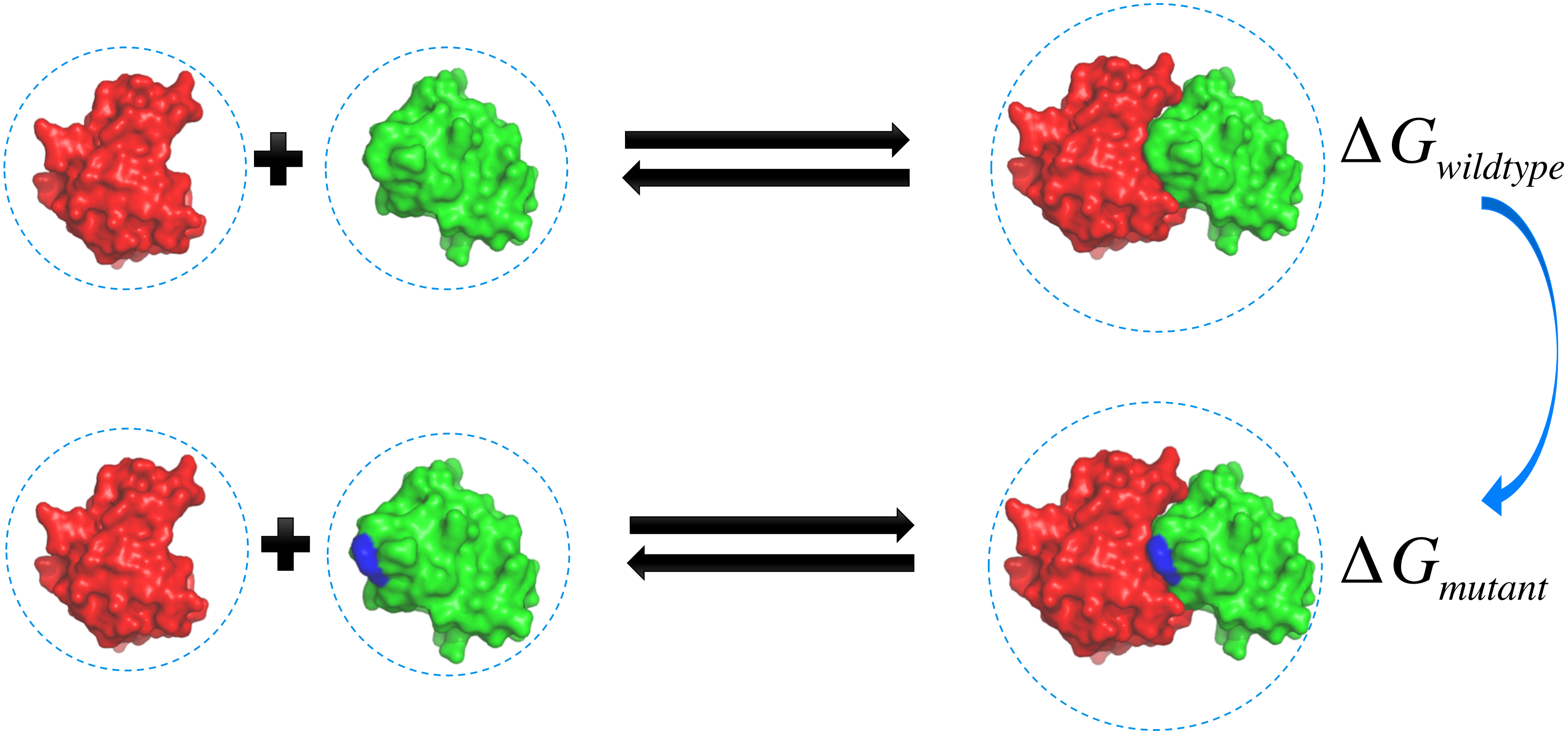
CC/PBSA (A Benedix et al., 2009), pred1 and pred2 (A Panchenko et al., 2014)

MD simulations + PBSA for solvation effect

FoldX **(R Guerois et al., 2002)**

10 force field based energy terms + rotamers for sidechain flexibility

Binding affinity change upon mutation (ΔΔG)



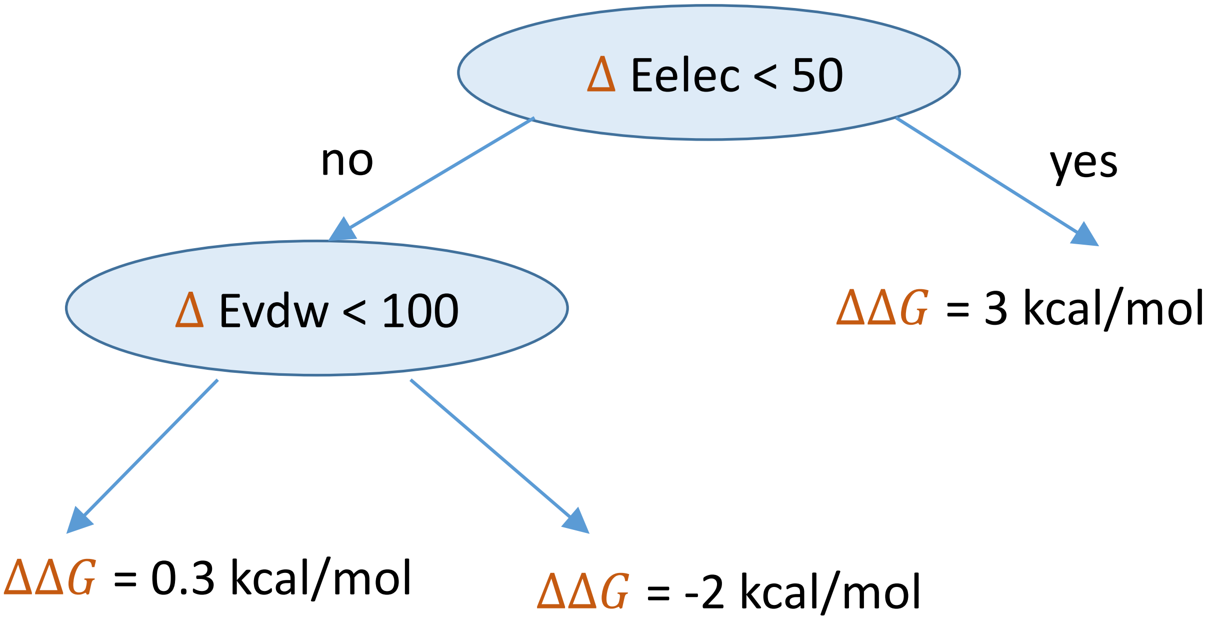
$$\Delta\Delta G = \Delta G_{mutant} - \Delta G_{wildtype}$$

iSEE uses random forest to predict $\Delta\Delta G$

iSEE aims to model:

Binding affinity change upon mutation ($\Delta\Delta G$) = f (features of the mutation site)

A random forest is a forest of decision trees.

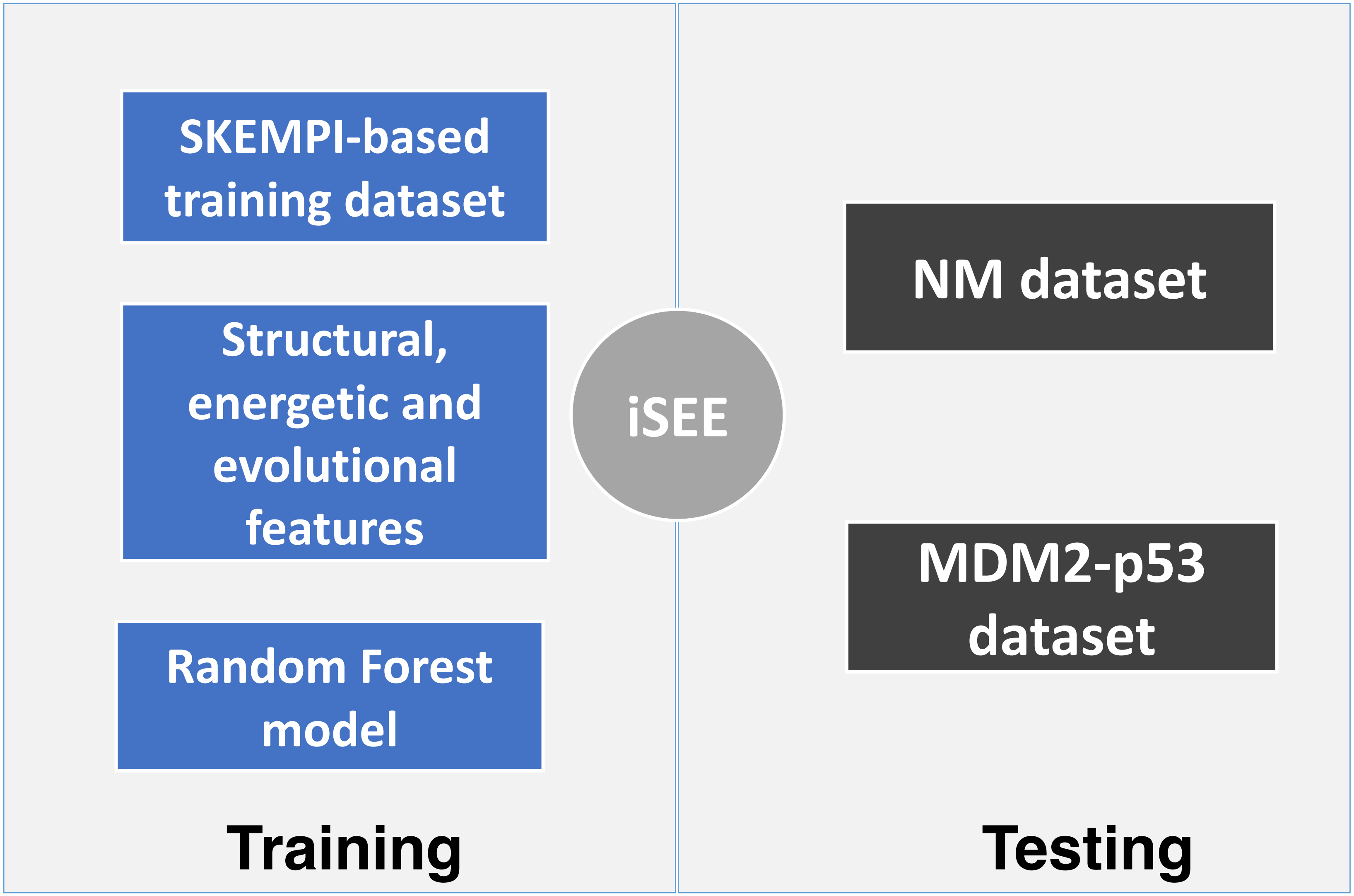

Δ Eelec < 50
no
yes
Δ Evdw < 100
$\Delta\Delta G$ = 3 kcal/mol
$\Delta\Delta G$ = 0.3 kcal/mol
$\Delta\Delta G$ = -2 kcal/mol

A conceptual decision tree.

INPUT: a feature vector of the mutation site

OUTPUT: predicted $\Delta\Delta G$

Methods



SKEMPI-based training dataset

Structural, energetic and evolutional features

Random Forest model

iSEE

NM dataset

MDM2-p53 dataset

Training

Testing

Training and Evaluation

10 times 10-fold cross validations are used for training.

Root Mean Square Error (RMSE):

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum_{1}^{n}(\Delta\Delta \boldsymbol{Gexp} \ - \Delta\Delta \boldsymbol{Gpred})^2}{\boldsymbol{n}}}$$

Pearson’s Correlation Coefficient (PCC):

$$\boldsymbol{PCC} = \frac{\boldsymbol{cov}(\Delta\Delta \boldsymbol{Gexp}, \Delta\Delta \boldsymbol{Gpred})}{\boldsymbol{\sigma}_{\Delta\Delta \boldsymbol{Gexp}} \cdot \boldsymbol{\sigma}_{\Delta\Delta \boldsymbol{Gpred}}}$$

Training and test datasets

- Wildtype protein-protein complexes with crystal structures and experimental ΔΔG values
- Only *single point mutations* in the *interface* of dimer complexes are considered.

Training dataset:

1102 mutations in 57 complexes from the SKEMPI database

Blind test datasets:

1. NM dataset

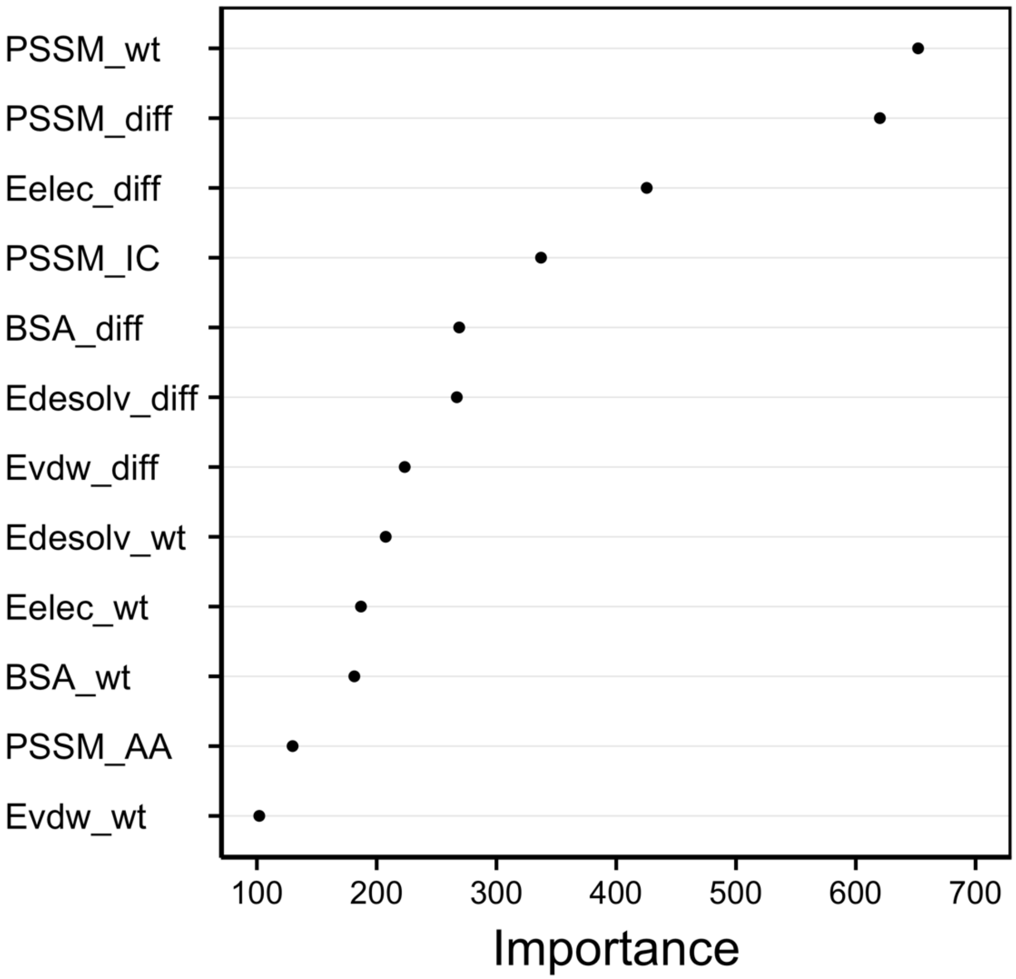
37 mutations in 2 complexes, 1BXI(18) and 1IAR(19).

2. MDM2-p53 complex dataset

33 mutations in the novel MDM2-p53 complex.

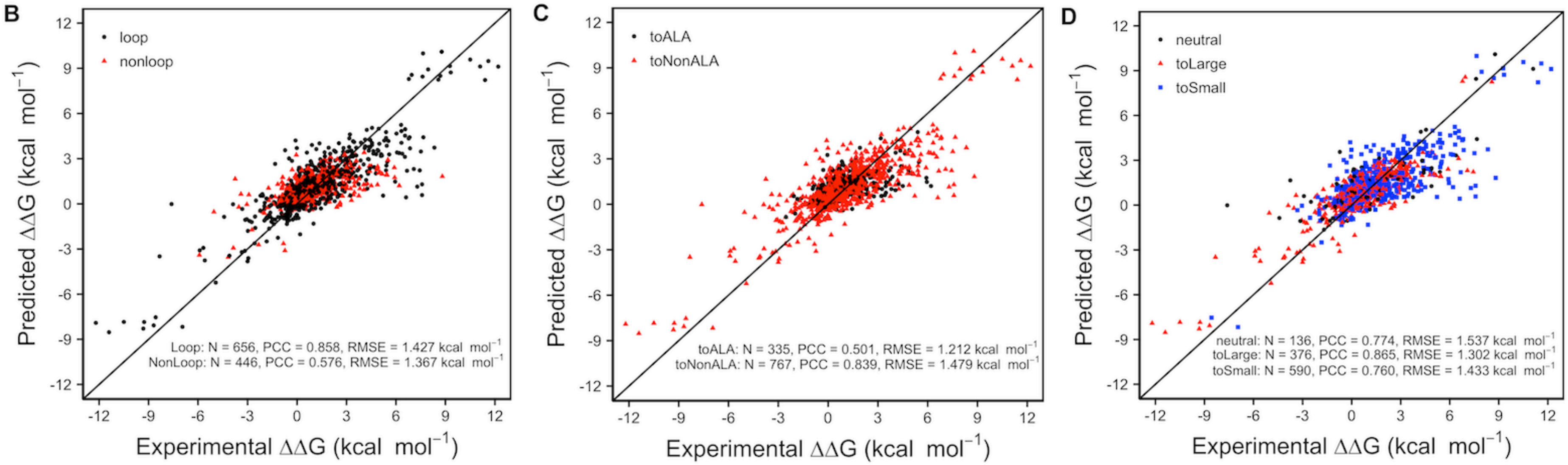
MDM2(16) and p53(17).

Feature importance



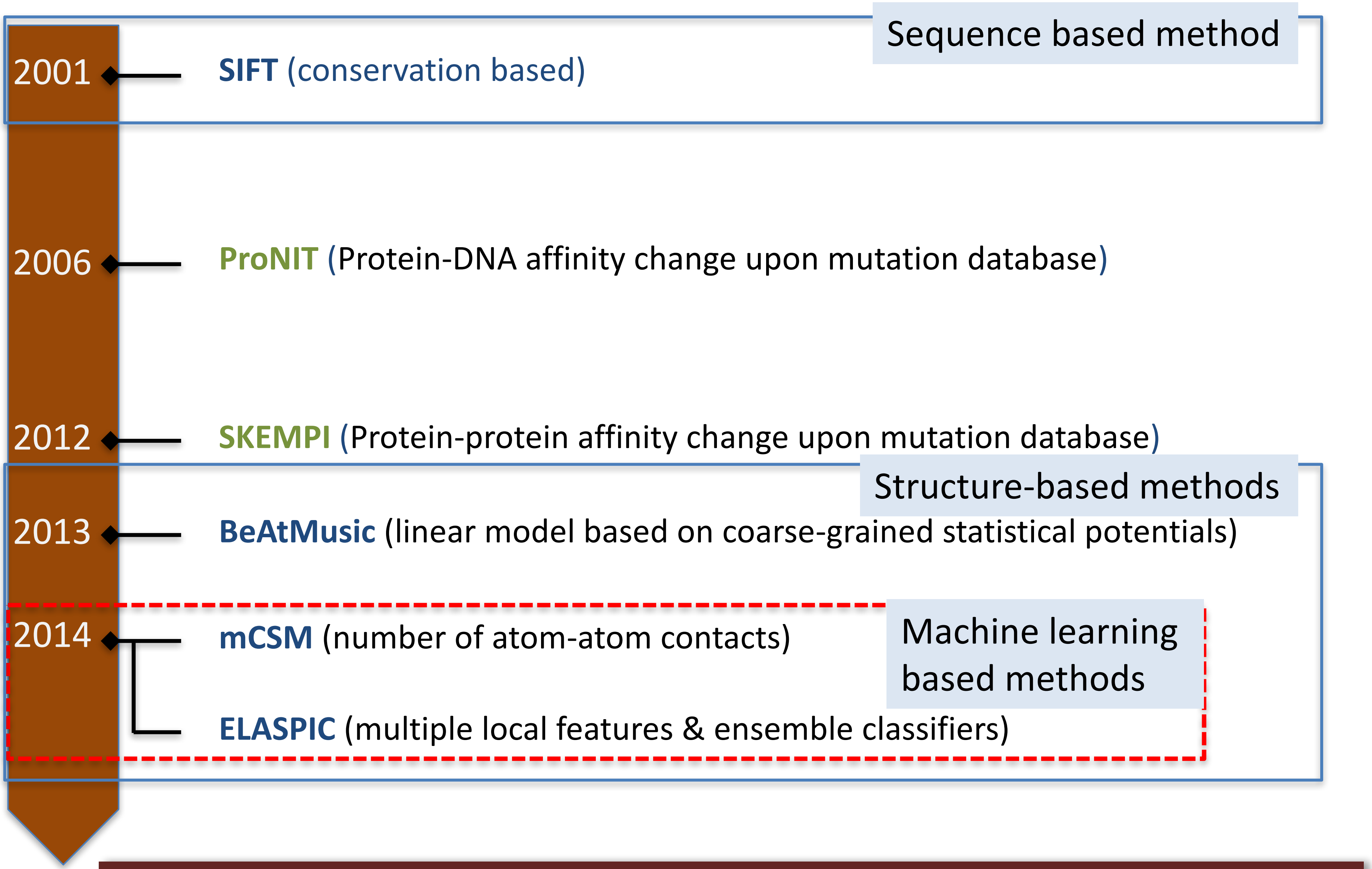
Evolutionary features ranked top.

iSEE performance on different types of mutations



iSEE highly and consistently performed on different types of mutations

Machine Learning based prediction methods



Sequence based method

2001 — **SIFT** (conservation based)

2006 — **ProNIT** (Protein-DNA affinity change upon mutation database)

2012 — **SKEMPI** (Protein-protein affinity change upon mutation database)

Structure-based methods

2013 — **BeAtMusic** (linear model based on coarse-grained statistical potentials)

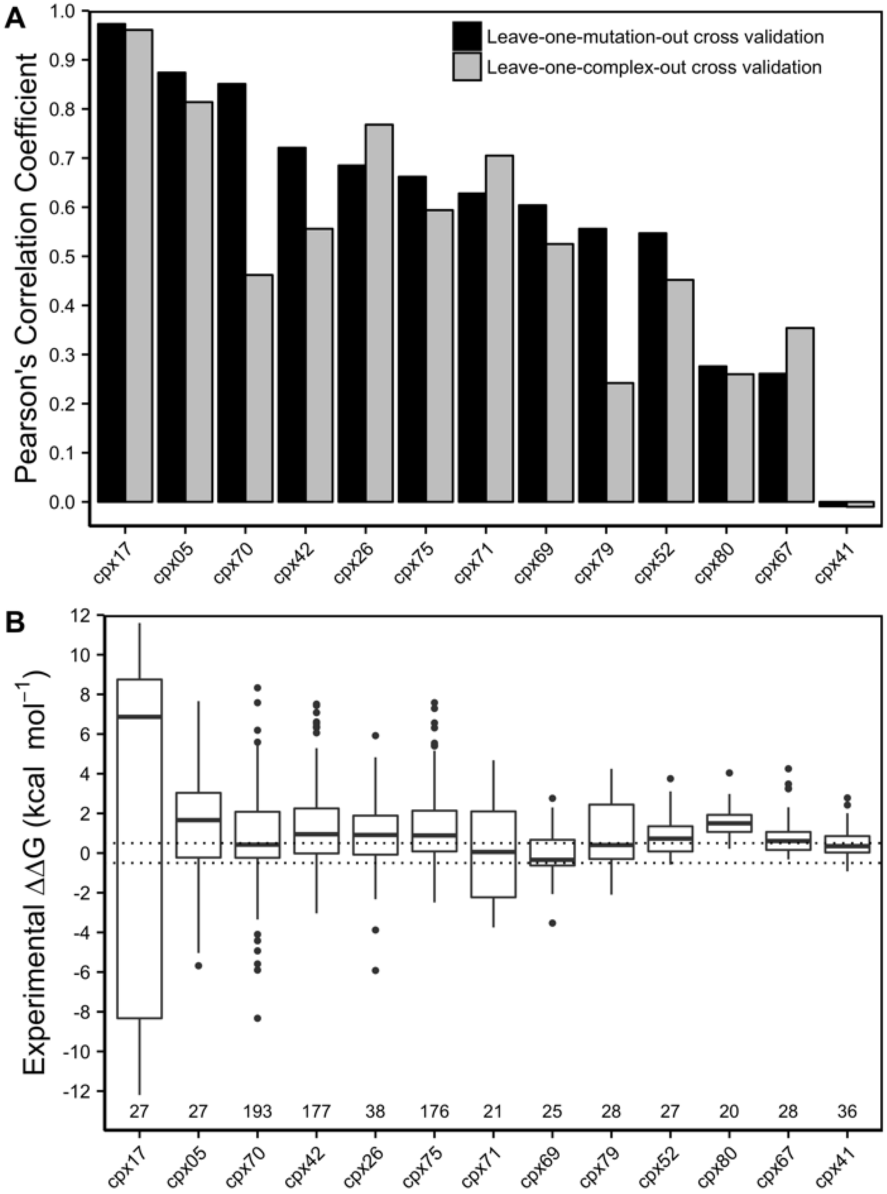
Machine learning based methods

2014 — **mCSM** (number of atom-atom contacts)

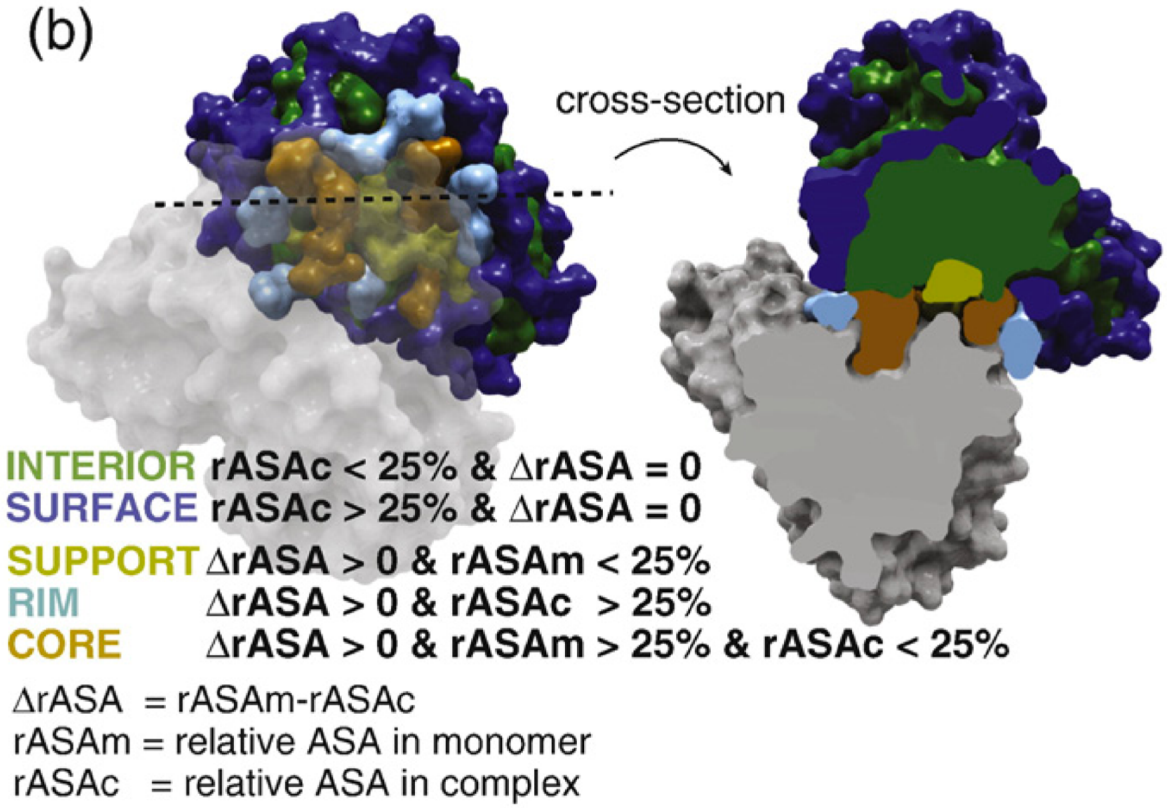
ELASPIC (multiple local features & ensemble classifiers)

Machine learning based methods are significantly better than other methods

Leave-one-type-out CV



Interface positions

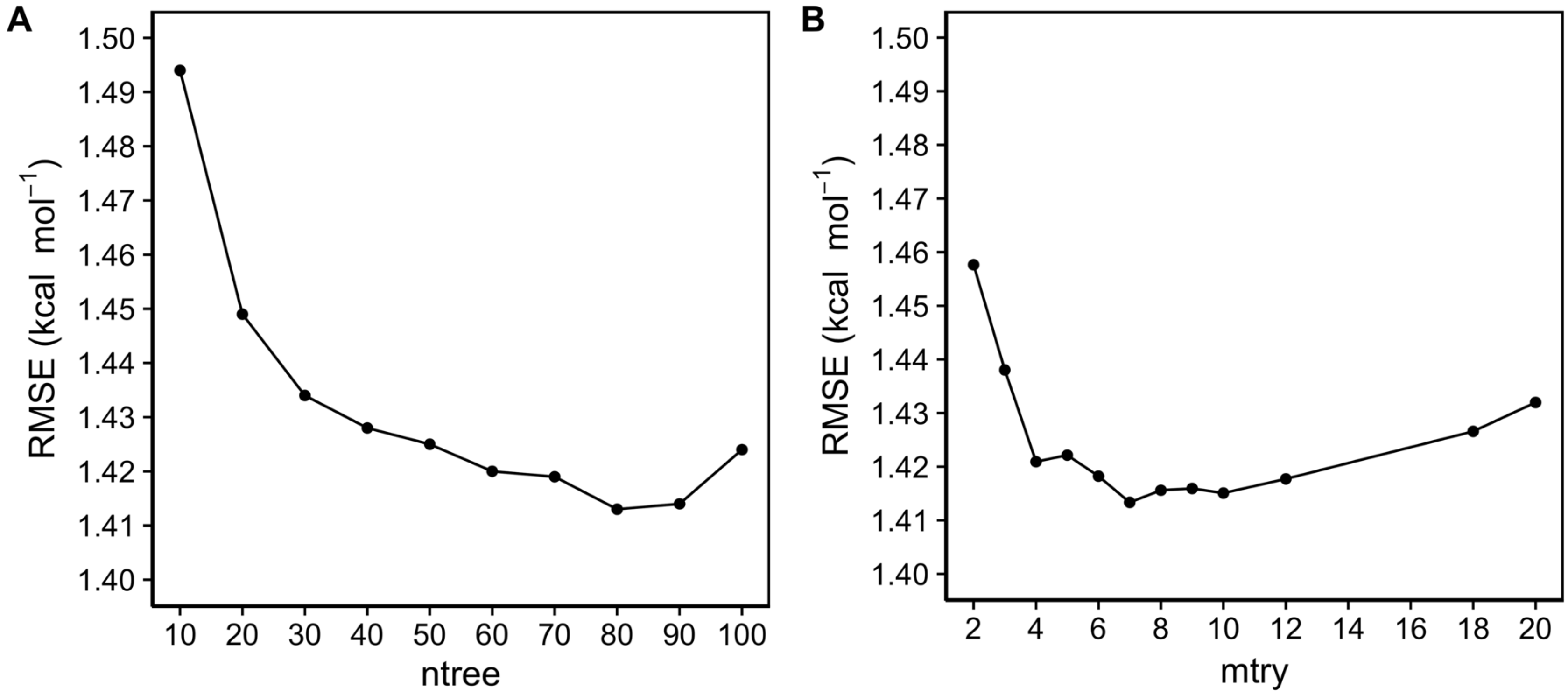


INTERIOR rASAc < 25% & ΔrASA = 0
SURFACE rASAc > 25% & ΔrASA = 0

SUPPORT ΔrASA > 0 & rASAm < 25%
RIM ΔrASA > 0 & rASAc > 25%
CORE ΔrASA > 0 & rASAm > 25% & rASAc < 25%

ΔrASA = rASAm-rASAc
rASAm = relative ASA in monomer
rASAc = relative ASA in complex

Optimization of RF models



Achieved highest performance at ntree=80 and mtry=7