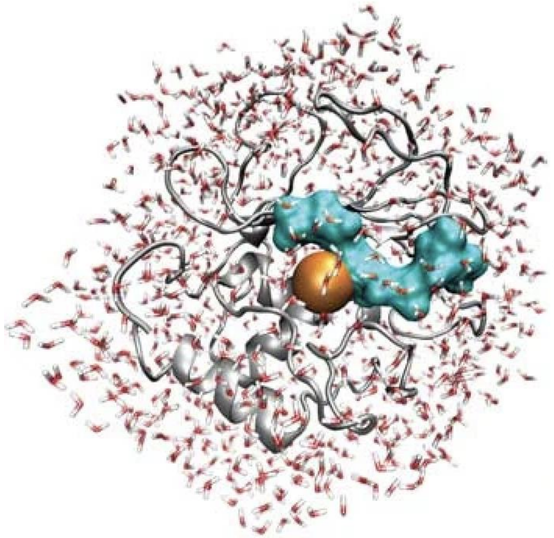
Leveraging quantum mechanical properties to predict solvent effects on large drug-like molecules

Mathias Hilfiker[1,2]
Leonardo Medrano Sandonas[3],Marco Klähn[2],Ola Engkvist[2], and Alexandre Tkathcenko[1]

[1] Department of Physics and Materials Science, University of Luxembourg, L-1511 Luxembourg City, Luxembourg.
[2] Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden
[3] Institute for Materials Science and Max Bergmann Center of Biomaterials, TU Dresden, 01062 Dresden, Germany.
mathias.hilfiker@astrazeneca.com, leonardo.medrano@tu-dresden.de

The water dilemma



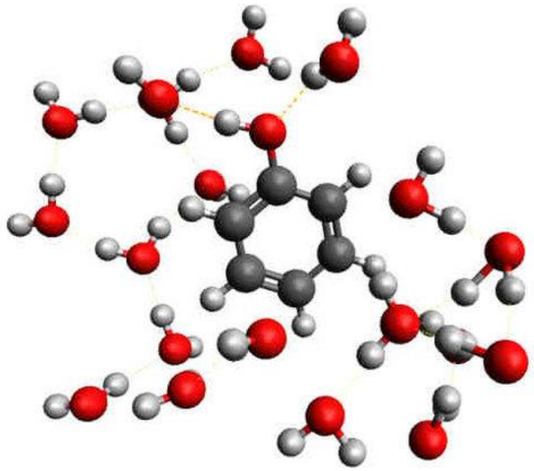
Source: Grossman, M., Born, B., Heyden, M. *et al.* N*at Struct Mol Biol* **18**, 1102–1108 (2011).

All the reactions that are interesting for pharmaceutical purposes happen in acqueous environments

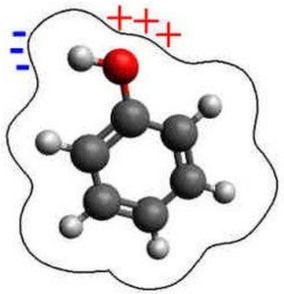
- Hydrogen bonding
- Van der Waals forces
- Structural reorganization

Describing solvation accurately is crucial for drug discovery

Explicit treatment



Implicit treatment



Adopting an implicit approach allows to reduce the required computations

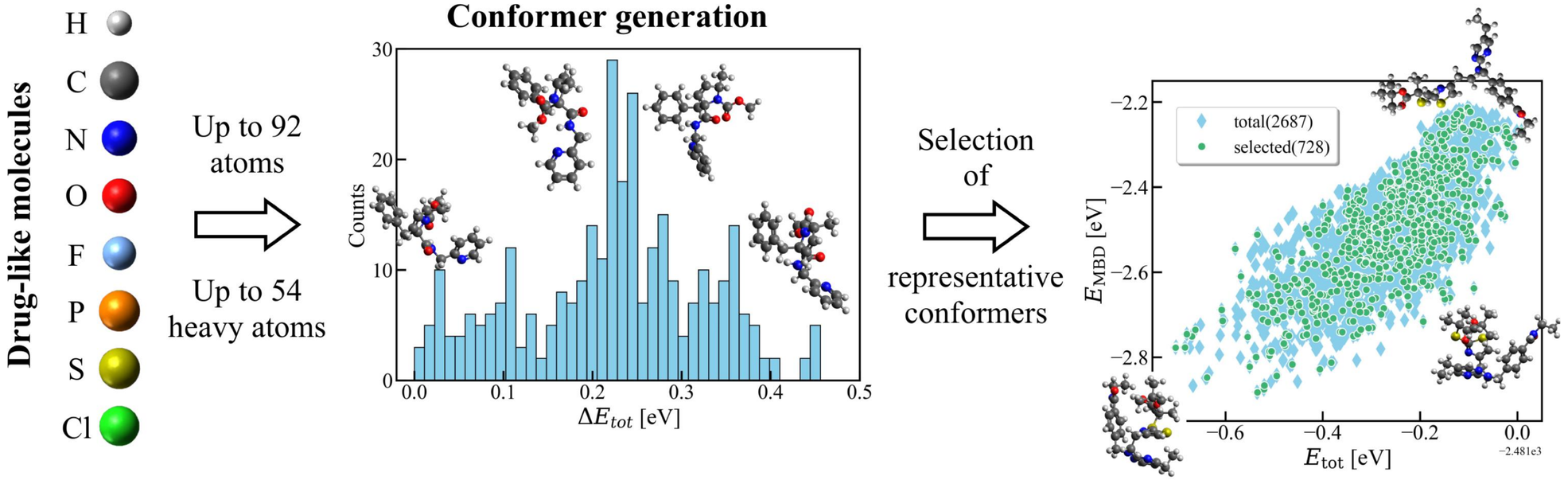
Comprehevnesive knoweldge of solvent effects is still lacking

Gaining general insights into the effect of solvation on structure/property and property/property relations for large drug-like molecules

Source: Wacławek, Stanisław. Ecological Chemistry and Engineering S, vol.28, no.1, 2021, pp.11-28.

Aquamarine (AQM) dataset[1]

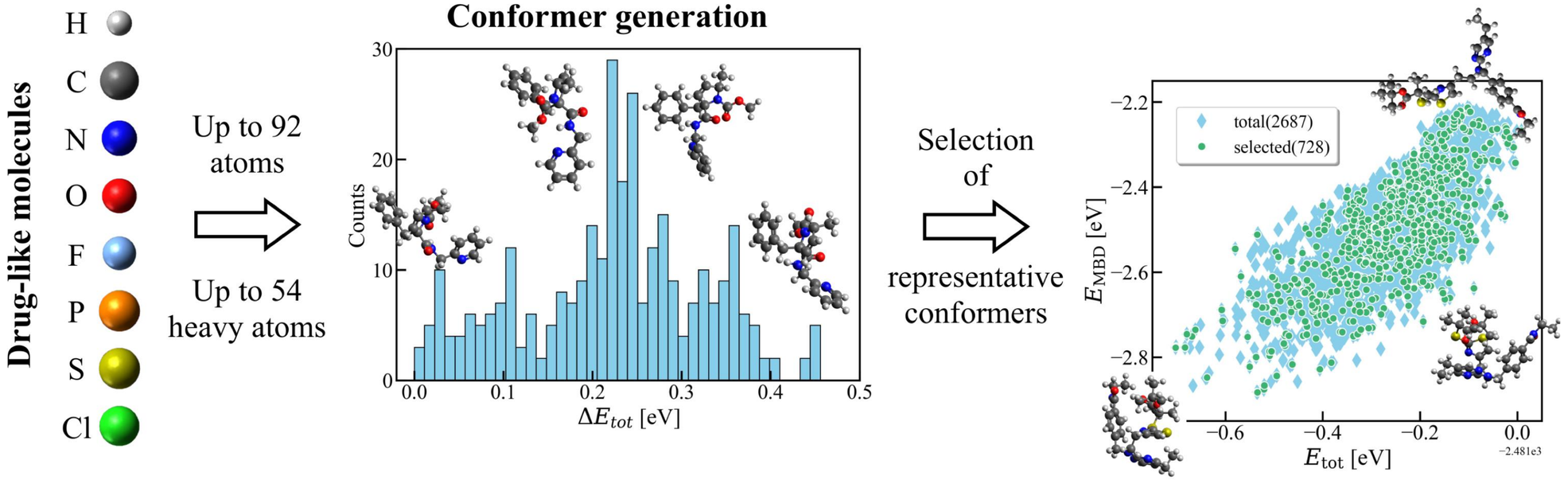
Dataset of QM properties for large drug-like molecules



[1] Medrano Sandonas, L. et al. *Scientific Data* 11.1 (2024): 742.

Aquamarine (AQM) dataset[1]

Dataset of QM properties for large drug-like molecules



[1] Medrano Sandonas, L. et al. *Scientific Data* 11.1 (2024): 742.

Aquamarine (AQM)

Dataset	Total structures	Total molecules	Max. total atoms	Max. heavy atoms	Elements	Geometry level of theory	Solvent model	Property level of theory	Solvent model	Total DFT properties
QMugs[38]	1,992,984	665,911	228	100	10	GFN2-xTB	—	ωB97X-D/def2-SVP	—	19
OE62[39]	30,876	30,876	174	92	16	PBE(tight)+TS	—	PBE0(tight)	MPE	3
BACE[40]	455,000	534	115	61	9	r2scan-3c/mTZVPP	C-PCM	r2scan-3c/mTZVPP	C-PCM	6
Amino acids[41]	1,300	26	96	39	5	Amber14 FF	explicit	ωB97M-D3(BJ) /def2-TZVPPD	—	10
AQM-gas	59,786	1,653	92	54	8	DFTB3+MBD	—	PBE0(tight)+MBD	—	36
AQM-sol	59,786	1,653	92	54	8	DFTB3+MBD	GBSA	PBE0(tight)+MBD	MPB	40

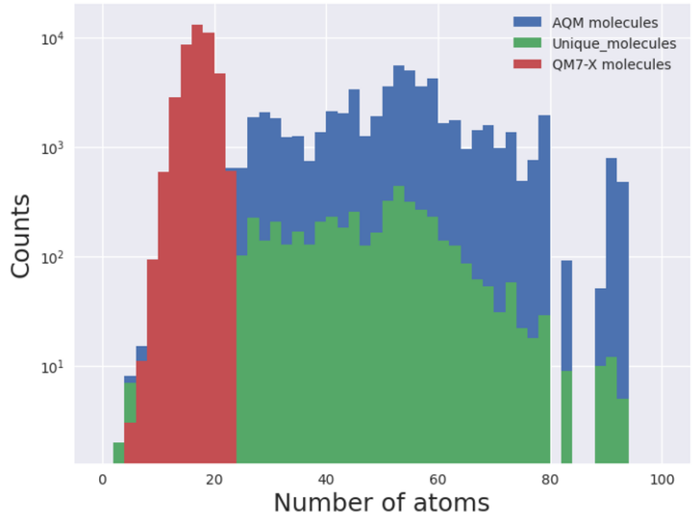
Table 1. Main characteristics of current quantum-mechanical datasets of large-sized molecular systems. We have selected datasets where molecule-solvent and/or van der Waals interactions have been considered during their generation procedure. The BACE and Amino acids sets have been extracted from the GEOM and SPICE collections.

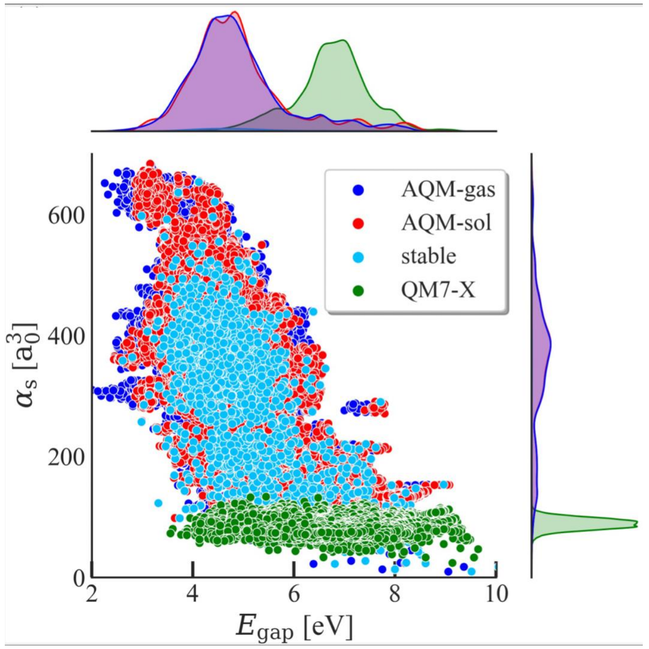
Key advancements:

- Exaustive selection of conformers;
- Large number of properties;
- Inclusion of long range interactions;
- Availability in both vacuum and water.

Comparison with QM7-X[2]

	QM7-X	AQM
Number of molecules:	40k	60k
Heavy atoms:	C,N,O,S,Cl	C,N,O,S,Cl,F,P
Number of properties	42	36 (40)
Chemical environment	No	Implicit water
Level of theory	PBE0+MBD	PBE0+MBD





[2] Hoja, J. et al. *Scientific data* 8.1 (2021): 43.

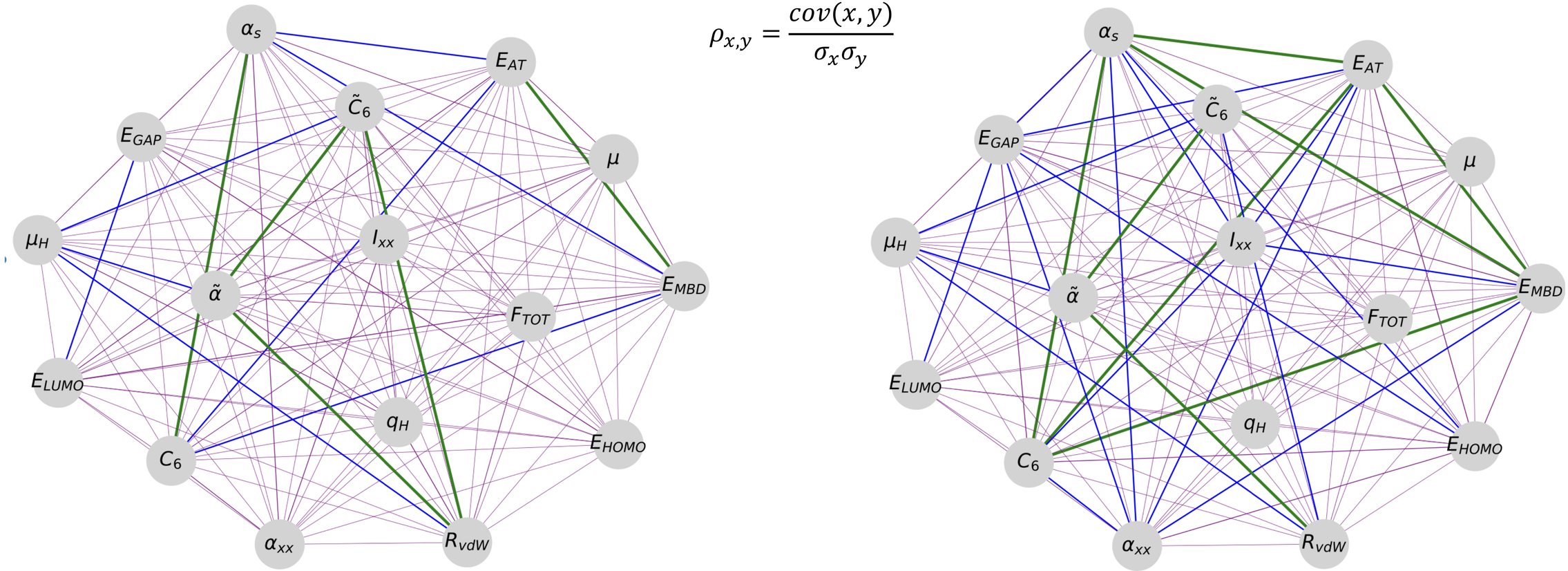
Correlation analysis: size effect

The property space of QM7-X is caracterized by a lack of correlations[3] (“Freedom of design”)
Does that sill hold true for large, solvated molecules?

$QM7 - X$

AQM_{gas}

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

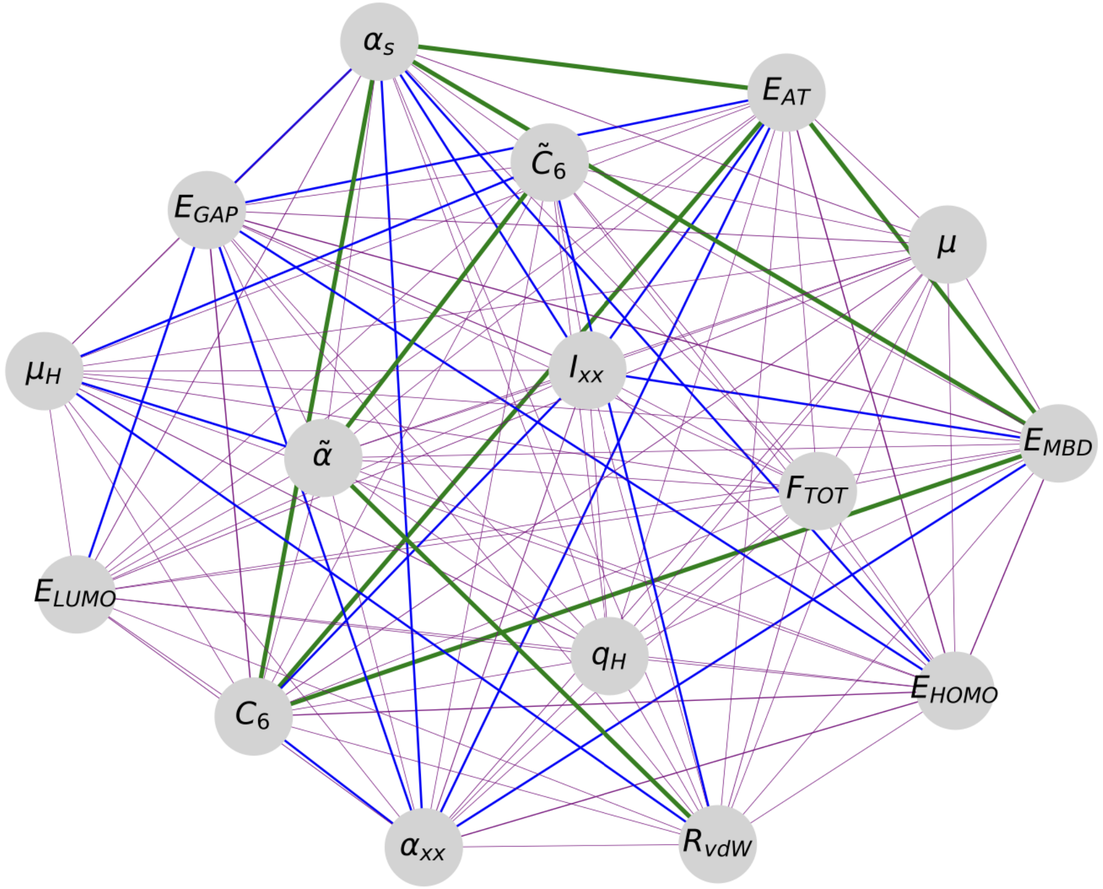


Moderately correlated couples:	8	18
Strongly correlated couples:	5	8

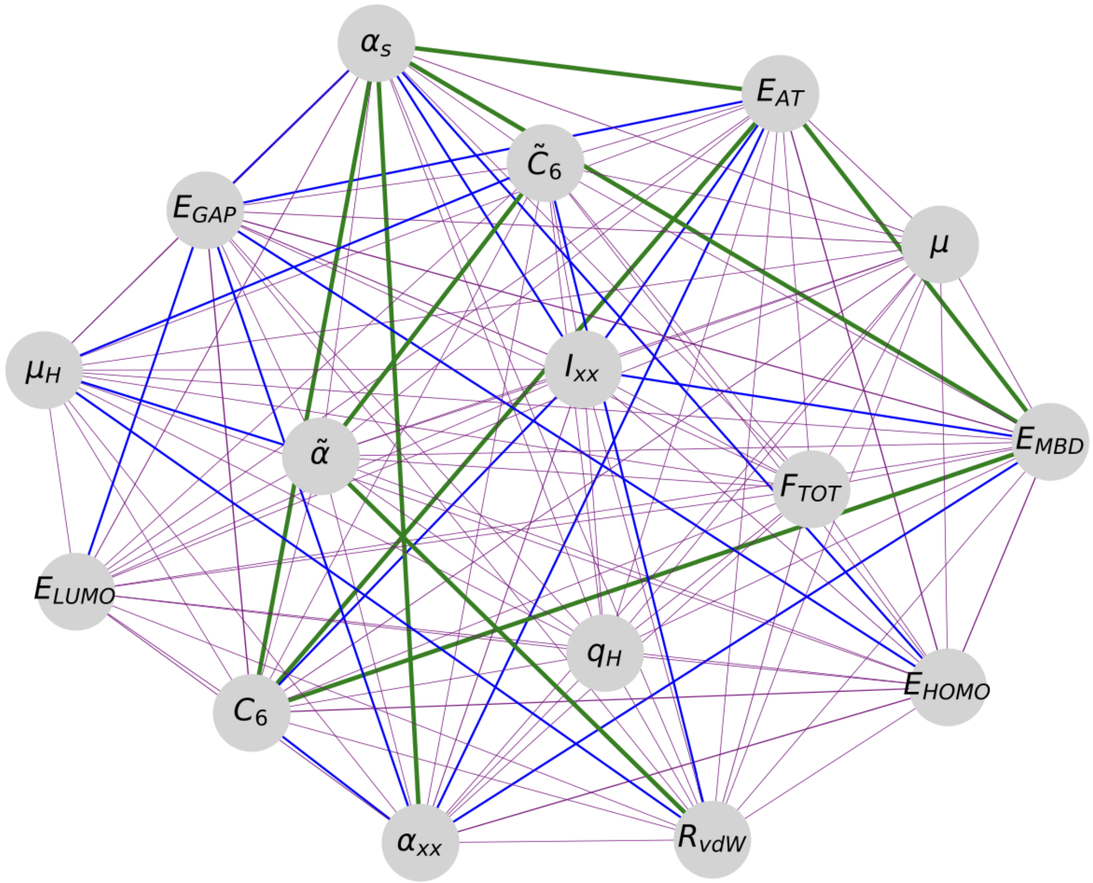
[3] Medrano Sandonas, L. et al. *Chemical Science* 14.39 (2023): 10702-10717.

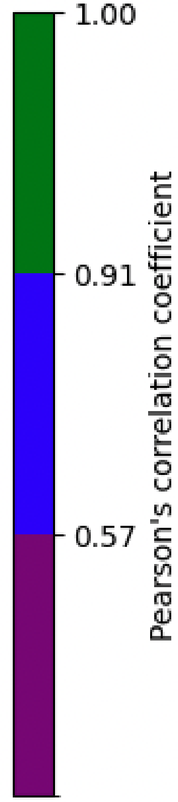
Correlation analysis: solvent effect

AQM_{gas}



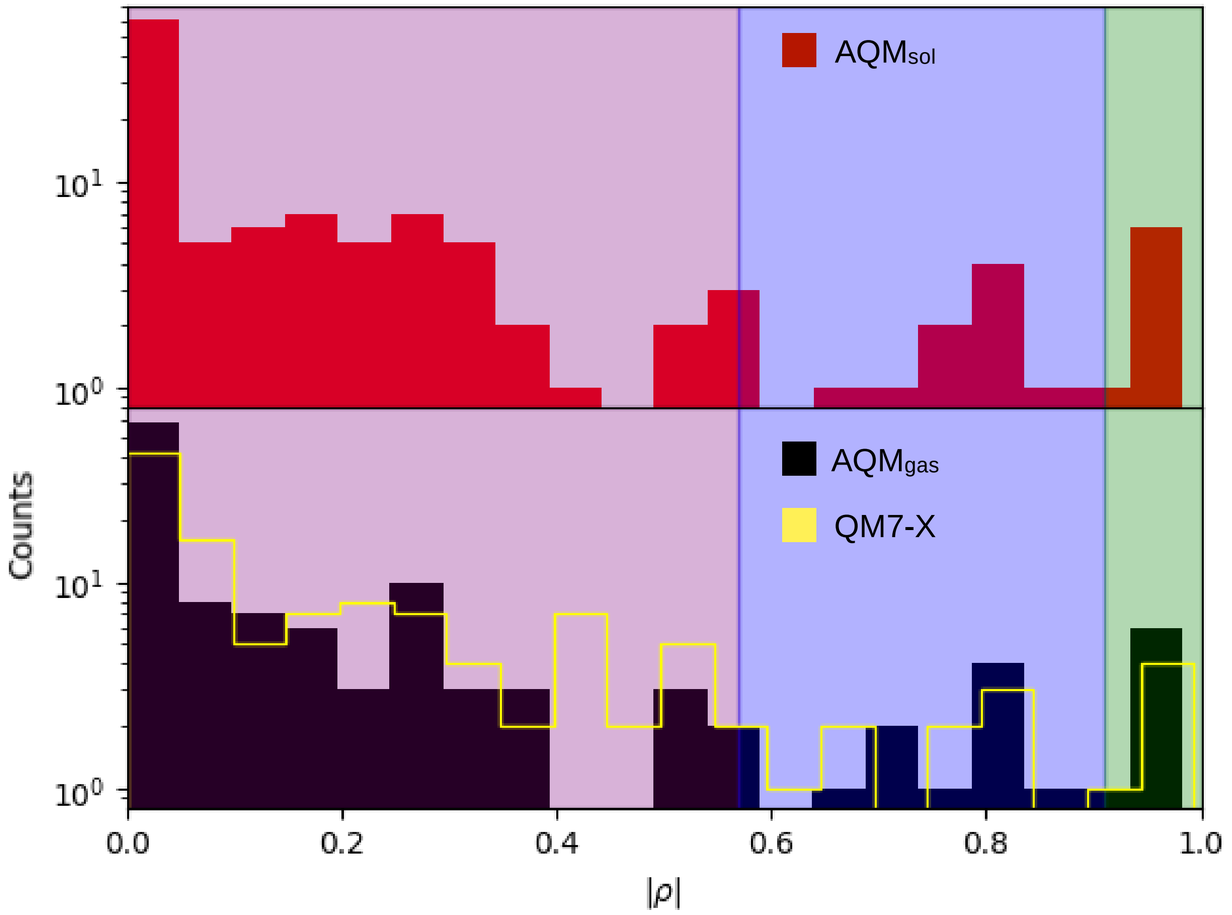
AQM_{sol}





Moderately correlated couples:	18	17
Strongly correlated couples:	8	9

Correlation analysis: conclusions



Weak correlation

Moderate correlation

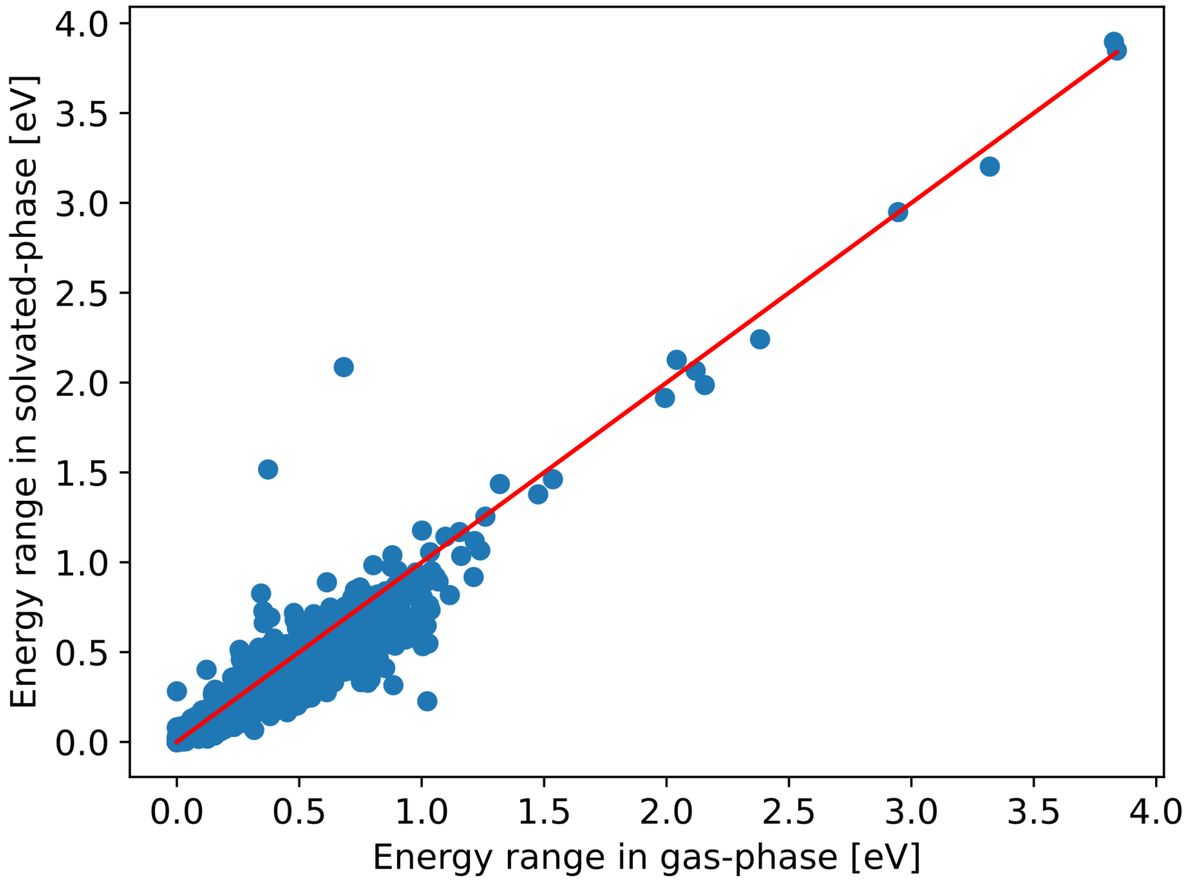
Strong correlation

- Properties are generally weakly correlated: “Freedom of design”retrieved
- Molecular size affects the correlation structure of the property space
- The acqueous environment affects the values of properties, but not the correlations between them

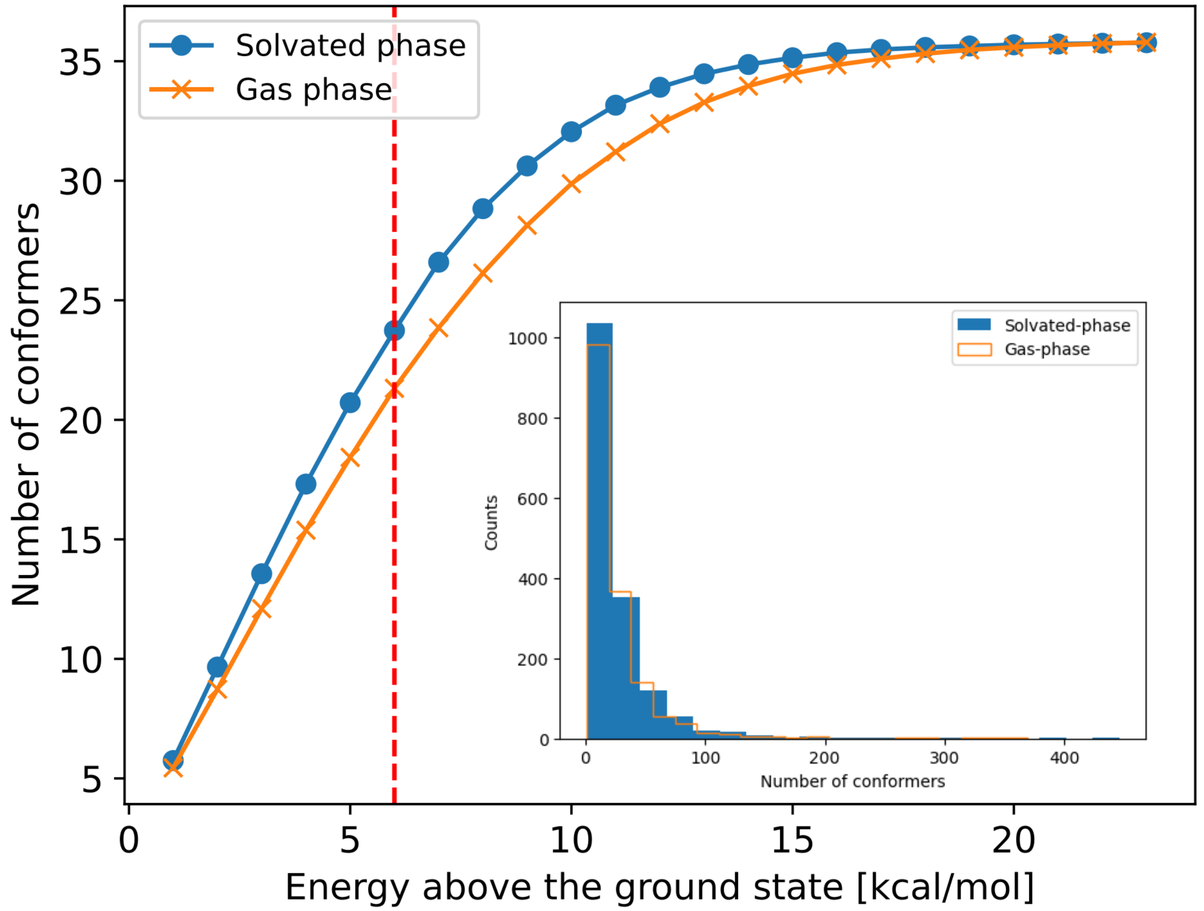
Energy analysis

How does solvation affect the distribution of conformers in the energy landscape?

$$E_{range,mol} = E_{max,mol} - E_{min,mol}$$



Conformers are compressed in a smaller energetic window in solvated phase

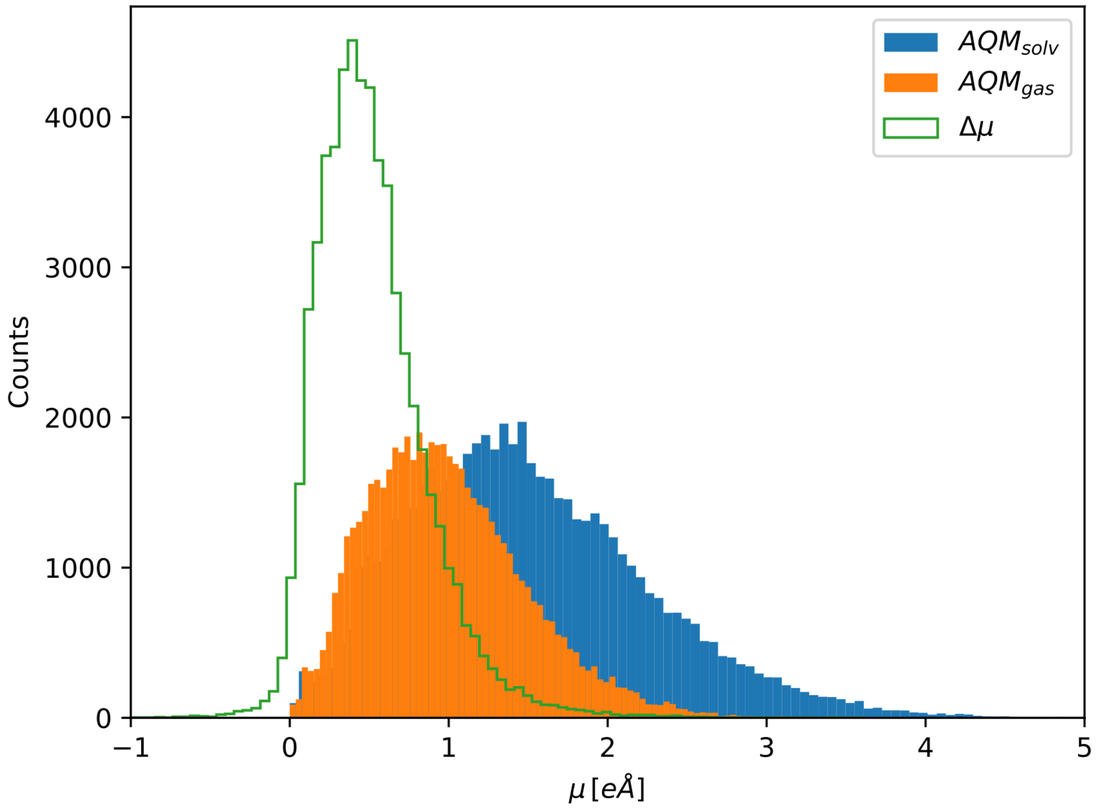


Solvation leads to a degeneration of energy levels

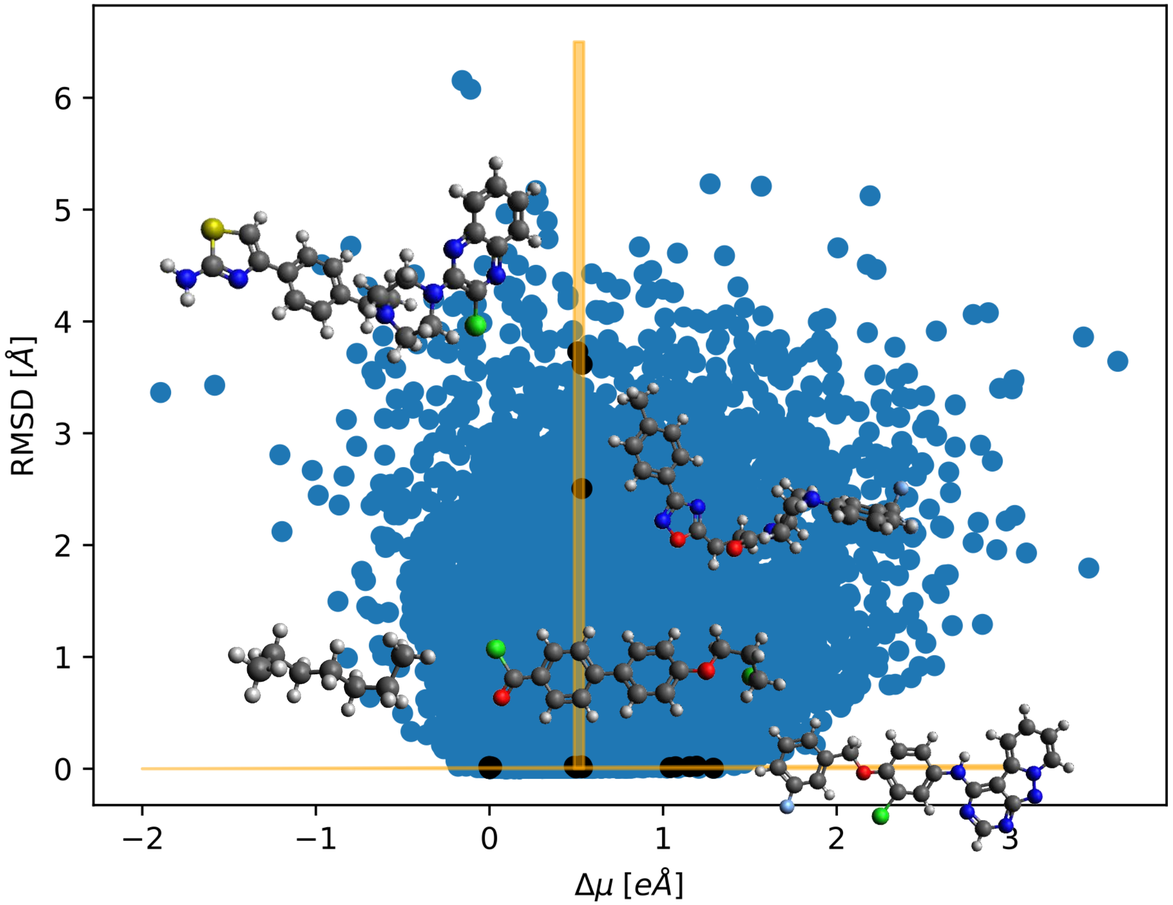
Structure-property relationship

How does the structural variation upon solvation affect properties?

$$\Delta\mu = \mu_{solv} - \mu_{gas}$$



$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$



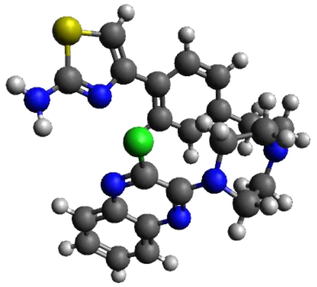
Freedom of design also in the Δ-space

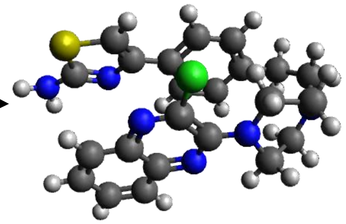
Δ-learning of solvent effects

Geometry optimization

Property calculation

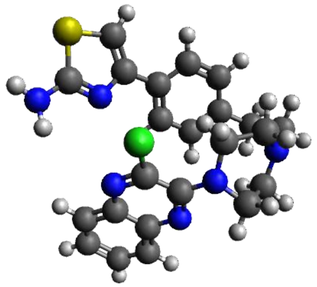
CREST Conformers

DFTB3+MBD →  → PBE0+MBD → P_{gas}

DFTB3+MBD + GBSA implicit solvent →  → PBE0+MBD + MPB implicit solvent → P_{sol}

$\Delta P = P_{sol} - P_{gas}$

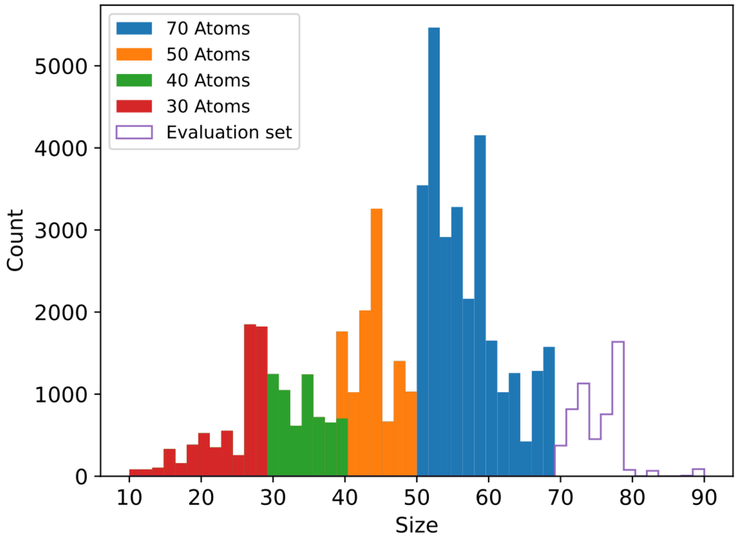
Geometry in gas phase



Equivariant Neural Network(s)[4][5]

ΔP — Δ-learning

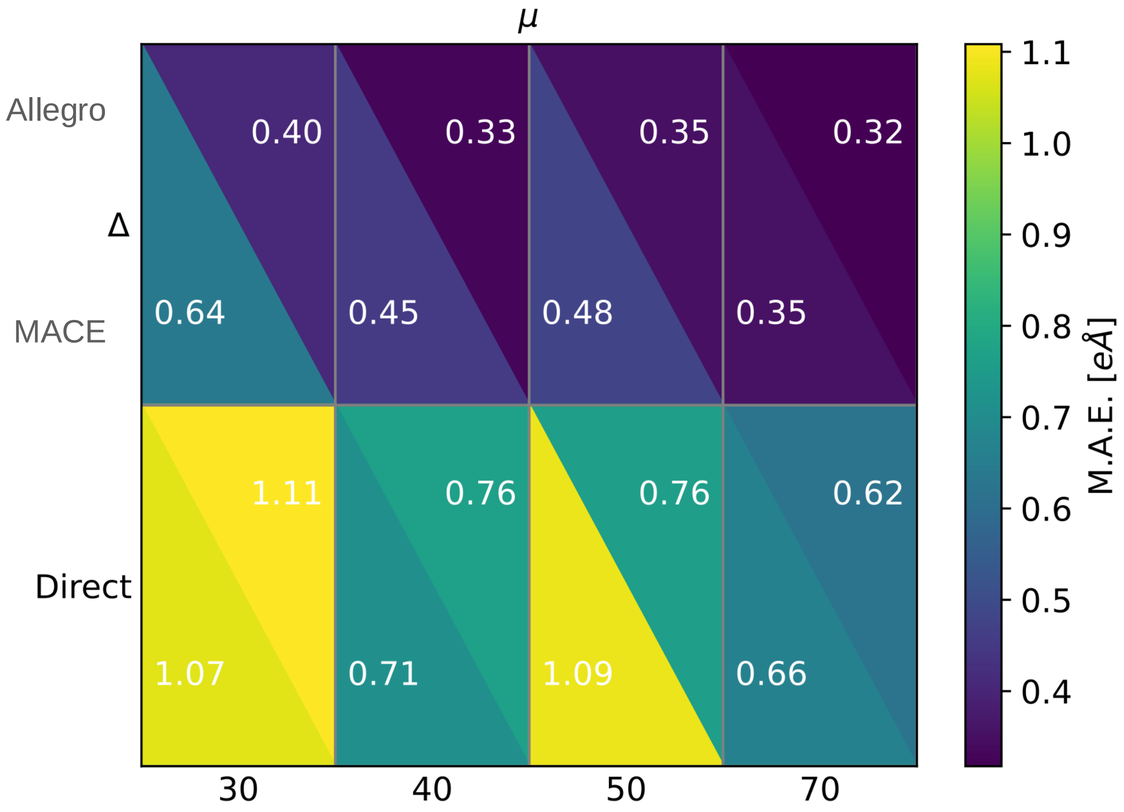
P_{sol} — Direct learning

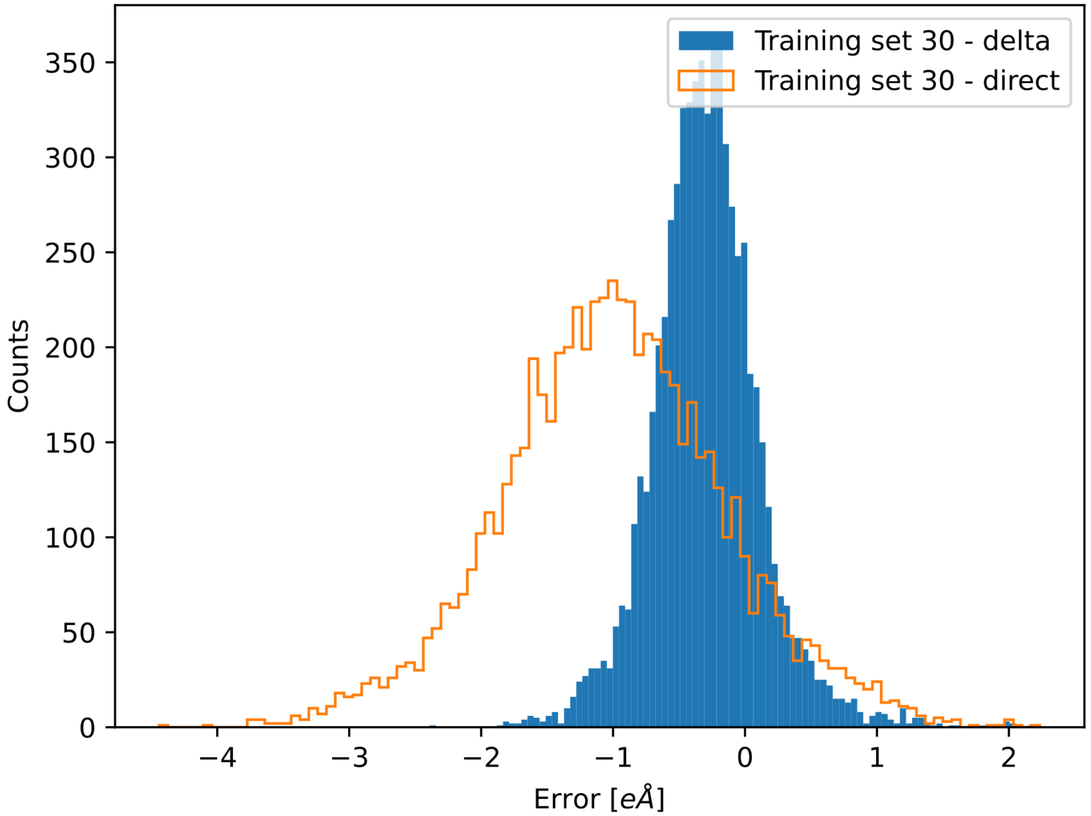


[4]: MACE (Batatia, I. et al. *Advances in Neural Information Processing Systems* 35 (2022): 11423-11436.)

[5]: Allegro (Musaelian, A. et al. *Nature Communications* 14.1 (2023): 579.)

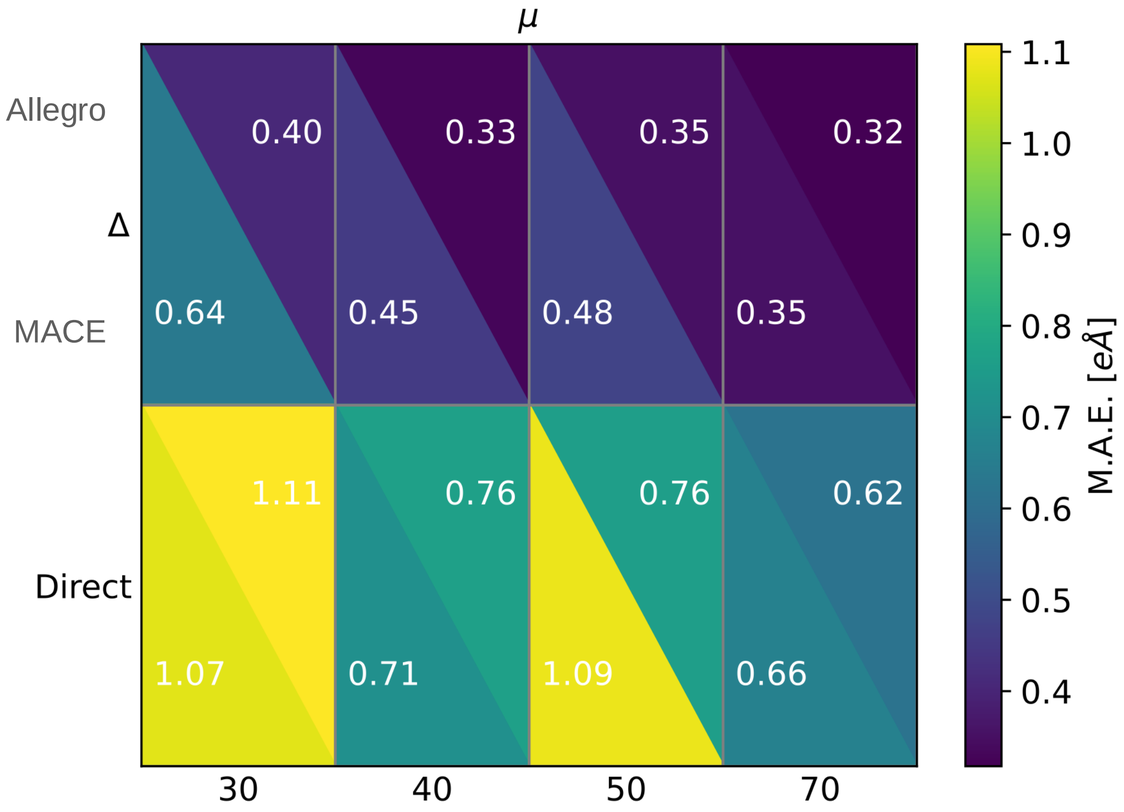
Δ-learning of solvated properties: dipole moment

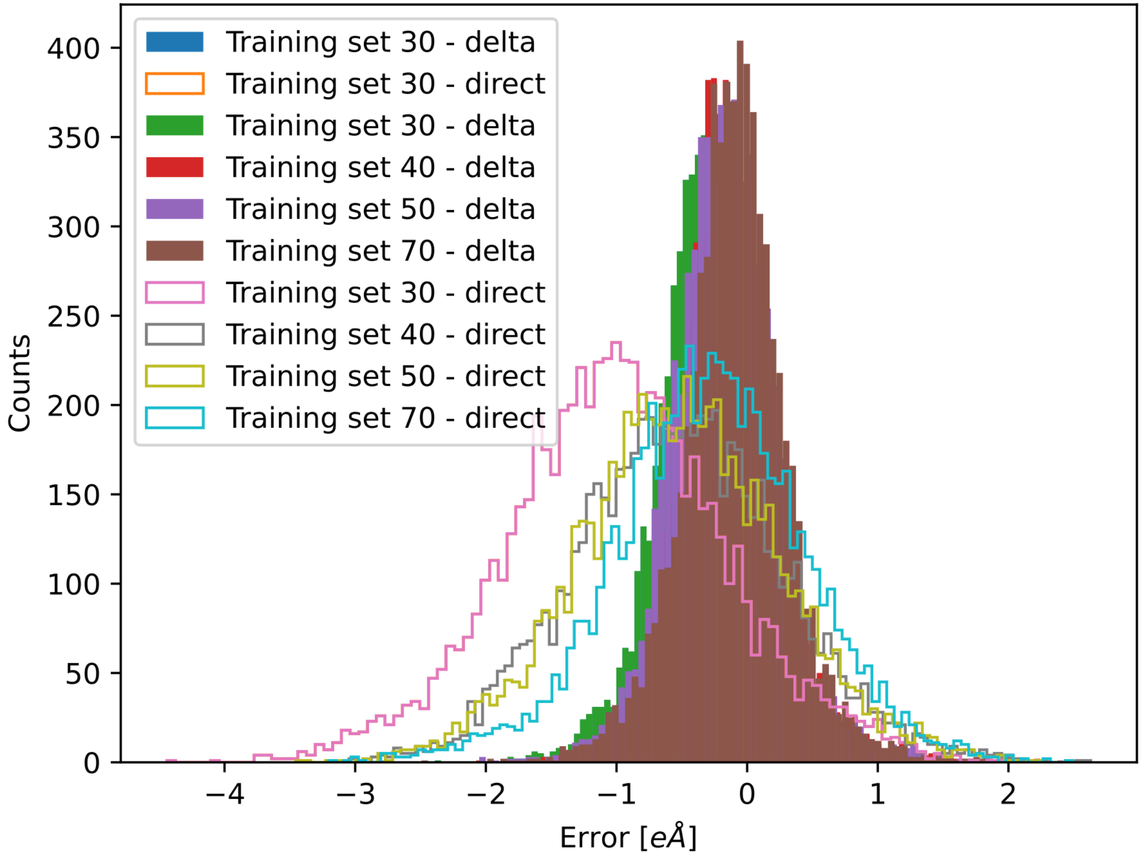




Δ-learning obtains less dispersed predictions

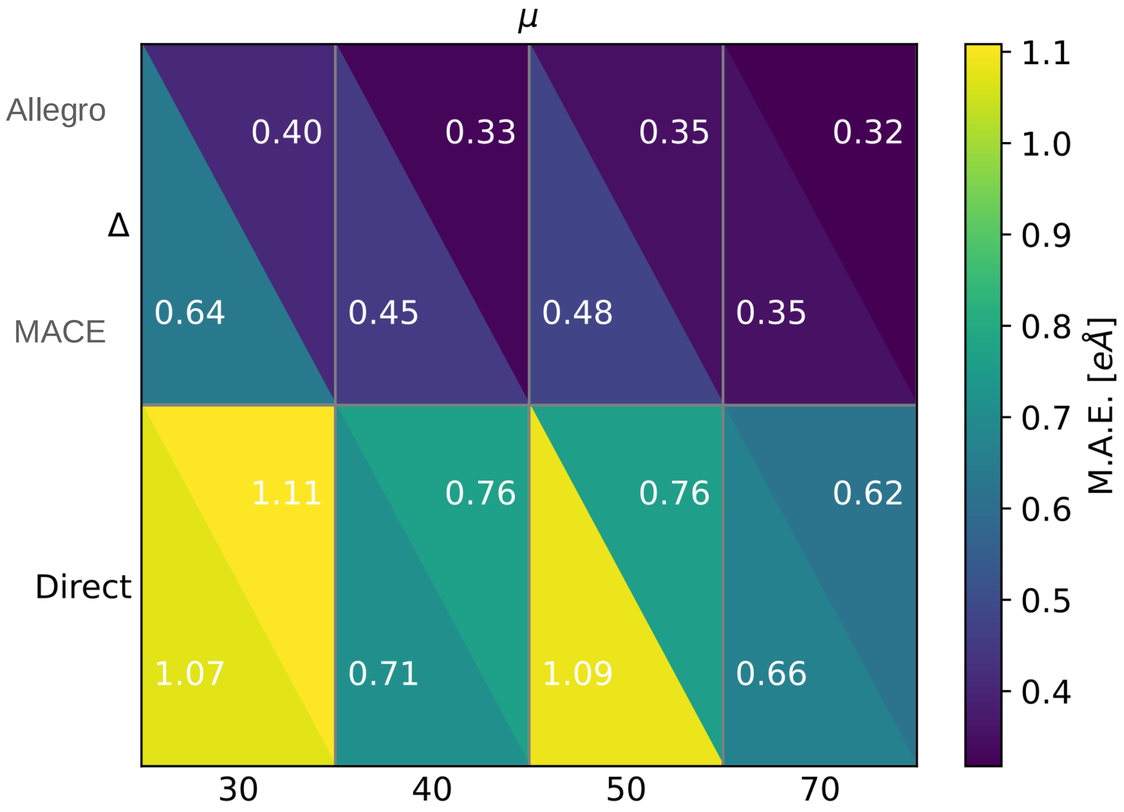
Δ-learning of solvated properties: dipole moment

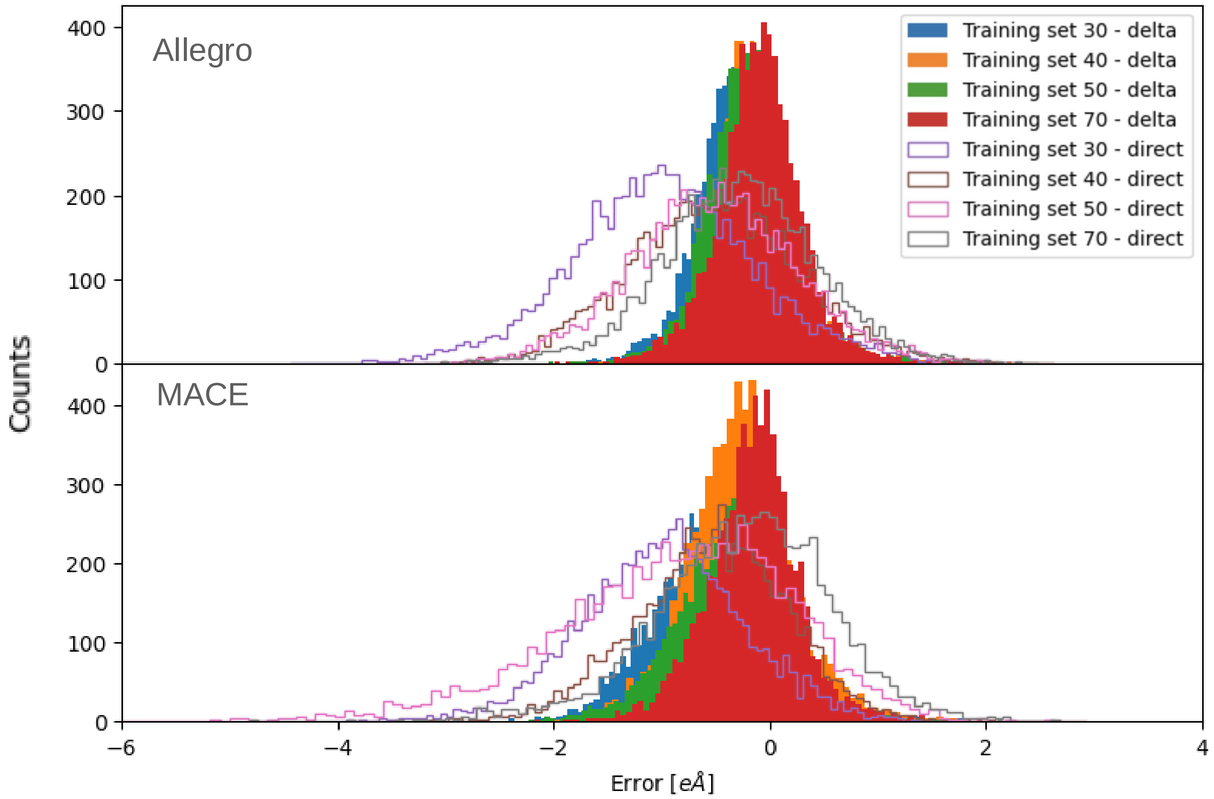




Δ-learning shows lower error and better scalability than direct learning

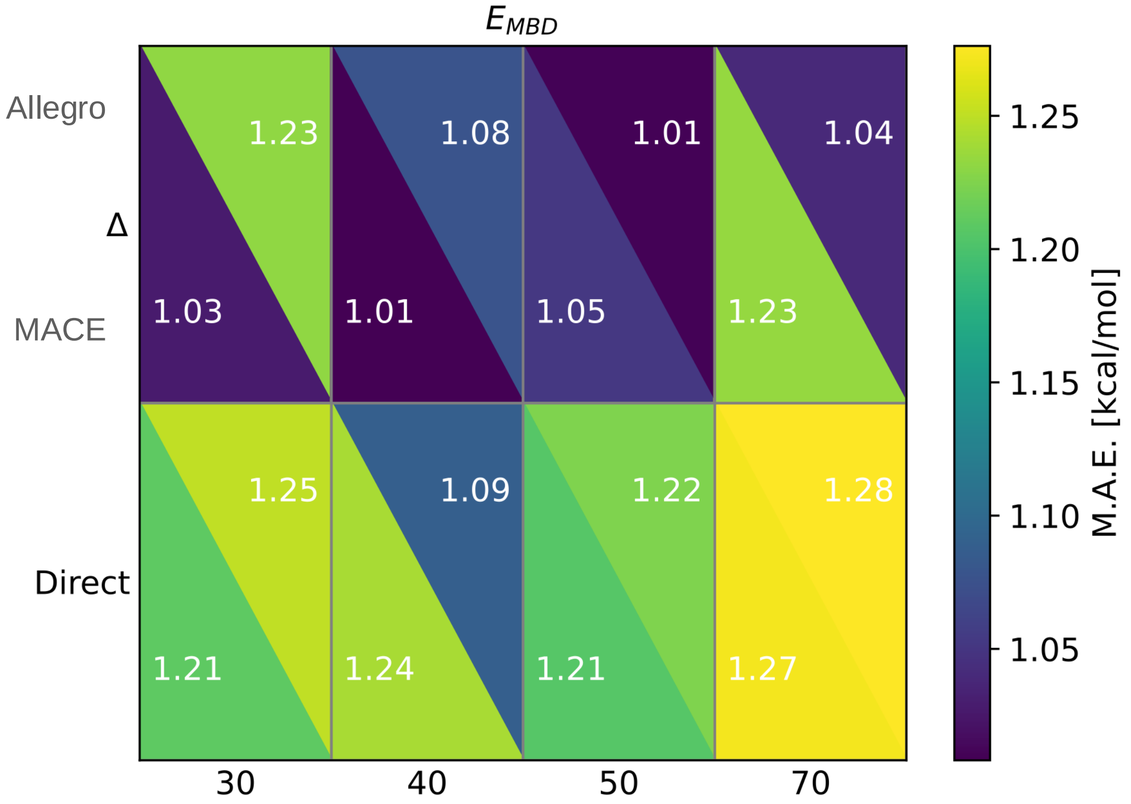
Δ-learning of solvated properties: dipole moment

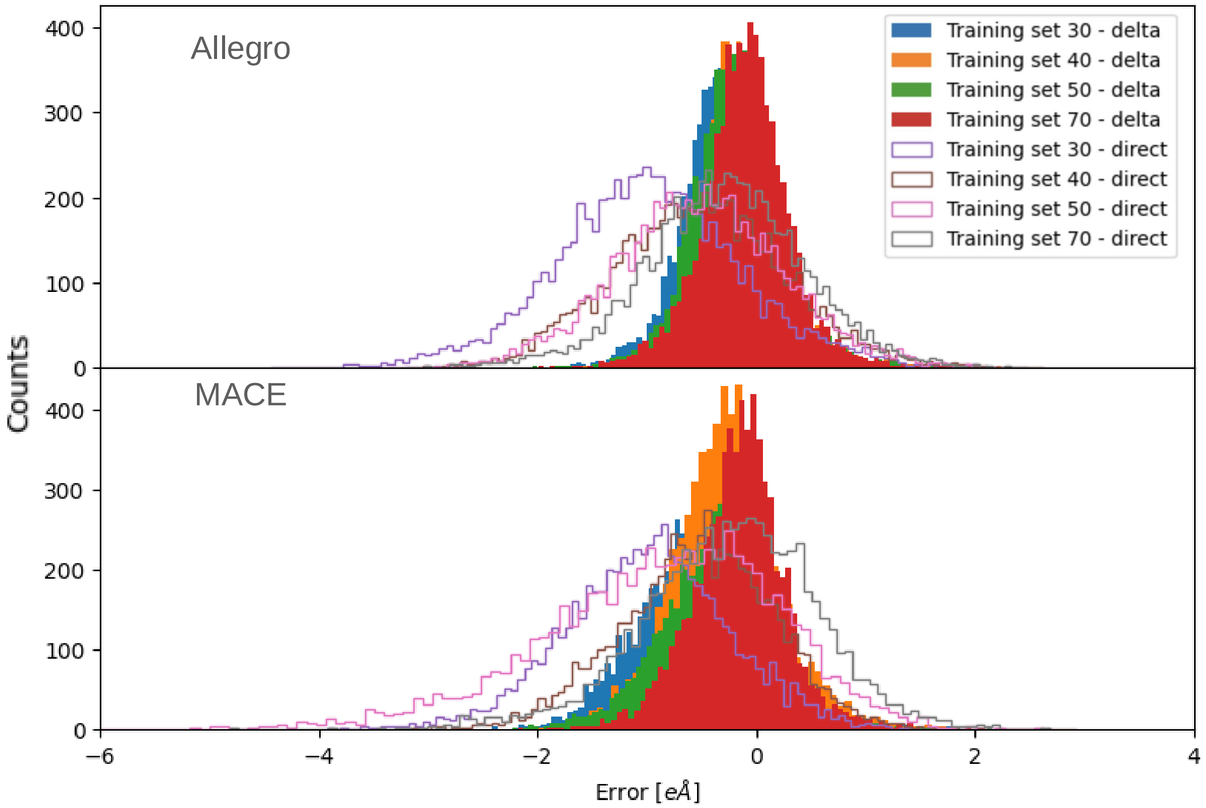




Δ-learning shows lower error and better scalability than direct learning

Δ-learning of solvated properties: Dispersion energy





Δ-learning achieves lower errors

Van der Waals interactions are ineherently long-range: 6Å cutoff is not enough!

Conclusions

Existing datasets can provide valuable insights into the effects that water has on large drug-like molecules

Solvation does not alter the correlations between properties	→	Correlations between properties may be transferable to the solvated phase
Solvation causes a degeneracy of energy levels	→	Necessity of a more rigourous molecular description for machine learning models
Structural and property variations are uncorrelated	→	Machine learning models should take into consideration the electronic features

Δ-learning shows a way to predict solvated phase properties from gas-phase molecules

Acknowledgements

TCP group

Dr. Leonardo Medrano Sandonas

Prof. Dr. Alexander Tkatchenko

AstraZeneca:

- Dr. Ola Engkvist
- Dr. Marco Klähn

AIDD consortium