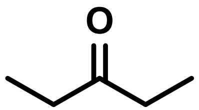
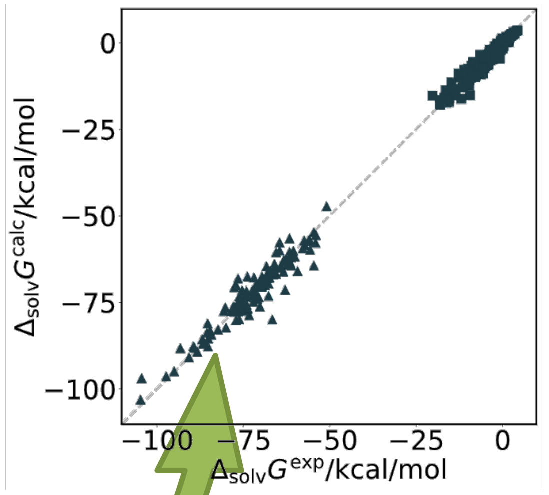
DoDSc Colloquium „Chemistry Meets Data Science“

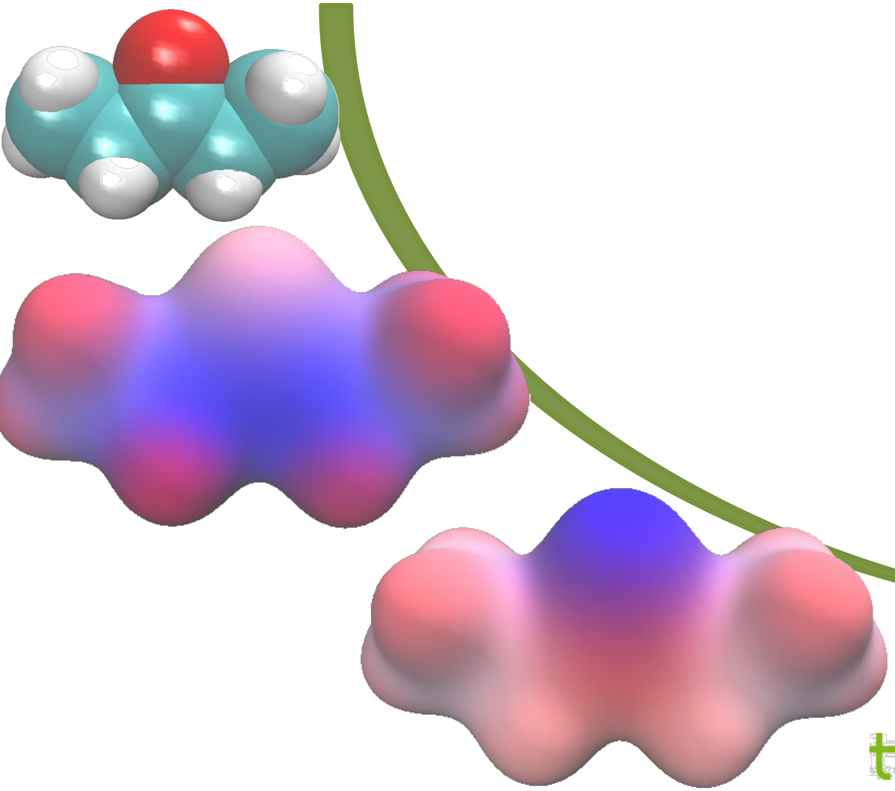


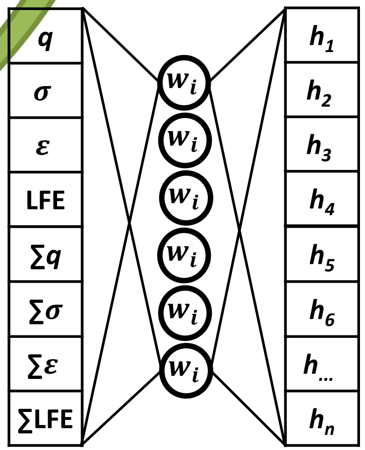
Machine Learning in Chemistry
Overview and Application Example

Stefan M. Kast
Christian Chodun

Department of Chemistry and Chemical Biology

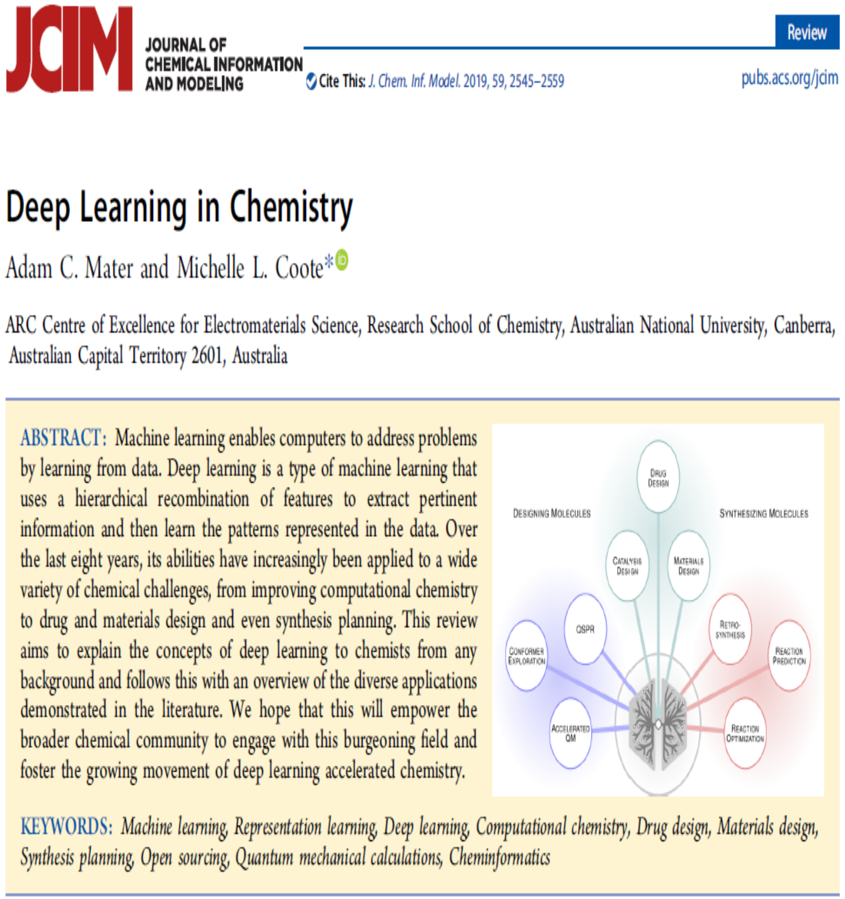






technische universität dortmund
fakultät für chemie und chemische biologie
RESOLV RUHR EXPLORES SOLVATION
CLUSTER OF EXCELLENCE - EXC 2033

Machine learning in chemistry



JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Review

Cite This: J. Chem. Inf. Model. 2019, 59, 2545−2559

pubs.acs.org/jcim

Deep Learning in Chemistry

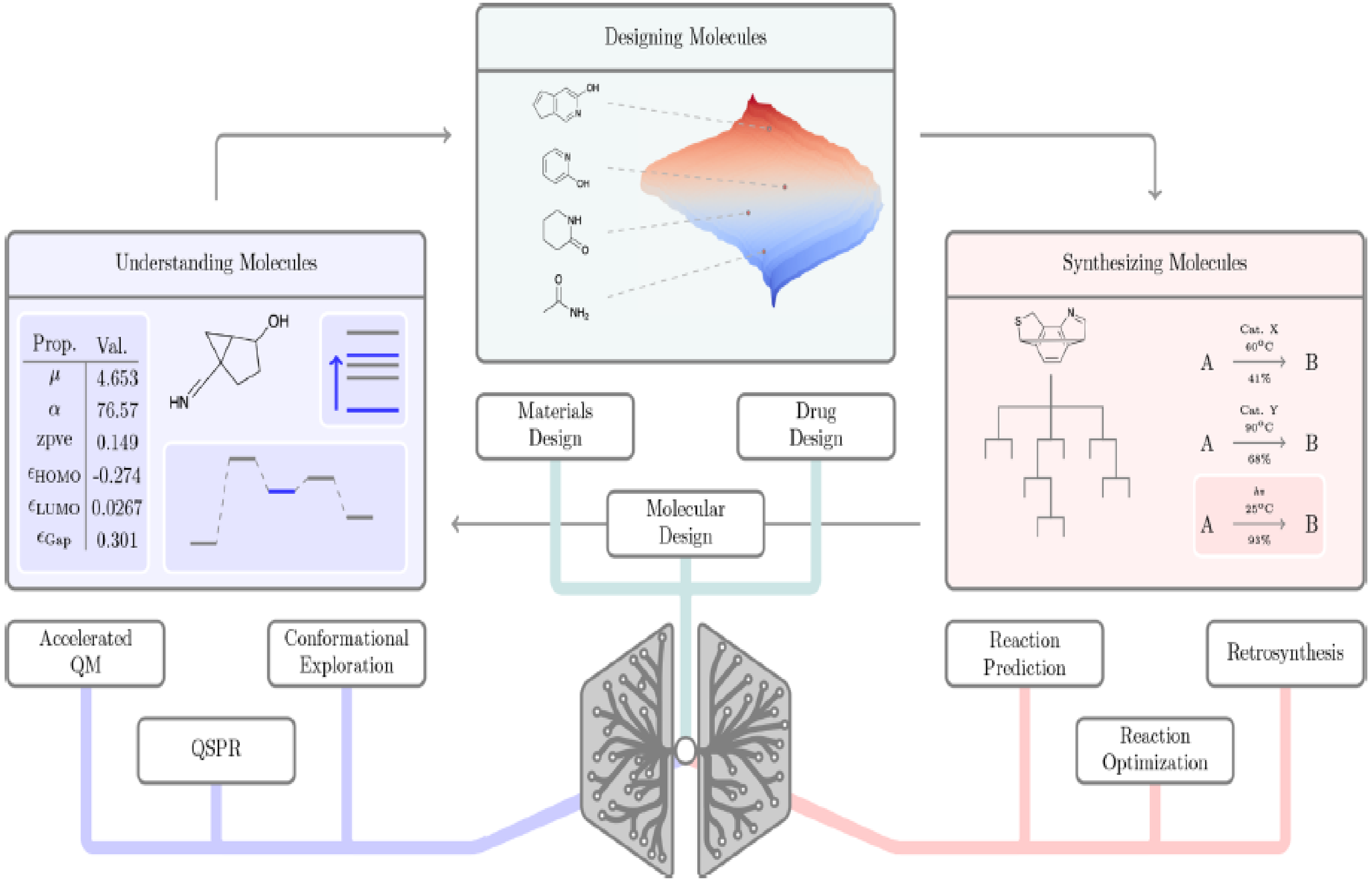
Adam C. Mater and Michelle L. Coote*

ARC Centre of Excellence for Electromaterials Science, Research School of Chemistry, Australian National University, Canberra, Australian Capital Territory 2601, Australia

ABSTRACT: Machine learning enables computers to address problems by learning from data. Deep learning is a type of machine learning that uses a hierarchical recombination of features to extract pertinent information and then learn the patterns represented in the data. Over the last eight years, its abilities have increasingly been applied to a wide variety of chemical challenges, from improving computational chemistry to drug and materials design and even synthesis planning. This review aims to explain the concepts of deep learning to chemists from any background and follows this with an overview of the diverse applications demonstrated in the literature. We hope that this will empower the broader chemical community to engage with this burgeoning field and foster the growing movement of deep learning accelerated chemistry.

KEYWORDS: Machine learning, Representation learning, Deep learning, Computational chemistry, Drug design, Materials design, Synthesis planning, Open sourcing, Quantum mechanical calculations, Cheminformatics

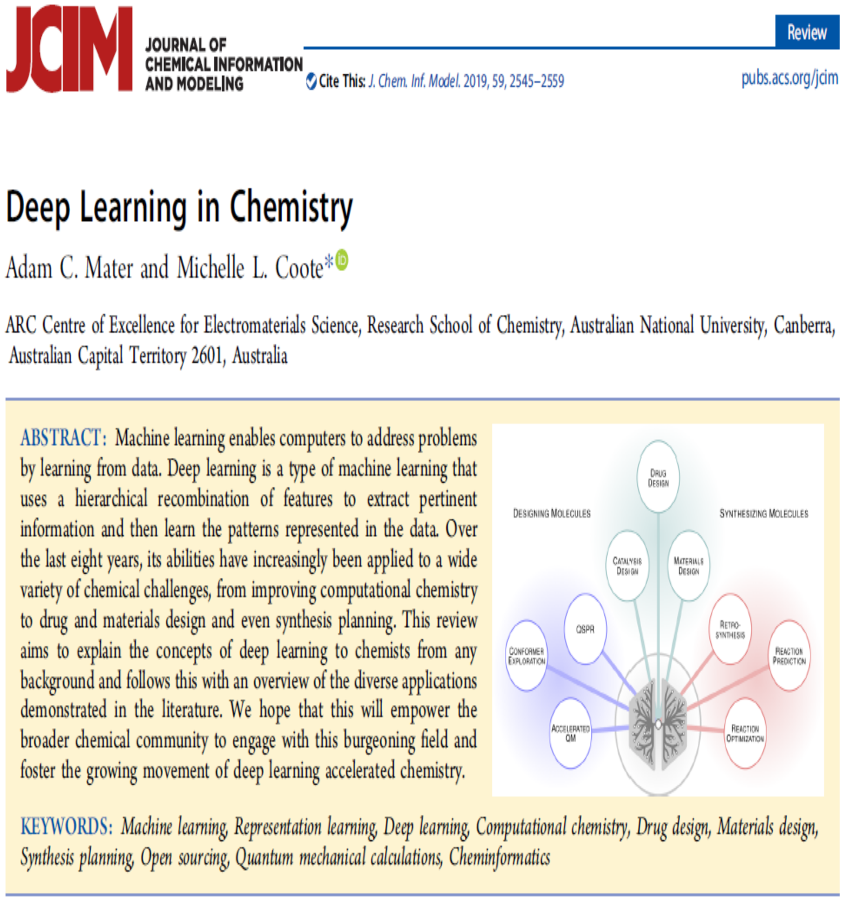
https://doi.org/10.1021/acs.jcim.9b00266



Prop.	Val.
μ	4.653
α	76.57
zpve	0.149
ϵ_{HOMO}	-0.274
ϵ_{LUMO}	0.0267
ϵ_{Gap}	0.301

Figure 5. Deep learning influence on the idealized chemical workflow. Illustrative examples of each task are shown in the dialogue boxed with arrows indicating the closed cycle that is contained within the framework. The property values in the blue panel were obtained from the QM9 data set for a randomly chosen molecule.[33]

Machine learning in chemistry



JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Review

Cite This: J. Chem. Inf. Model. 2019, 59, 2545−2559

pubs.acs.org/jcim

Deep Learning in Chemistry

Adam C. Mater and Michelle L. Coote*

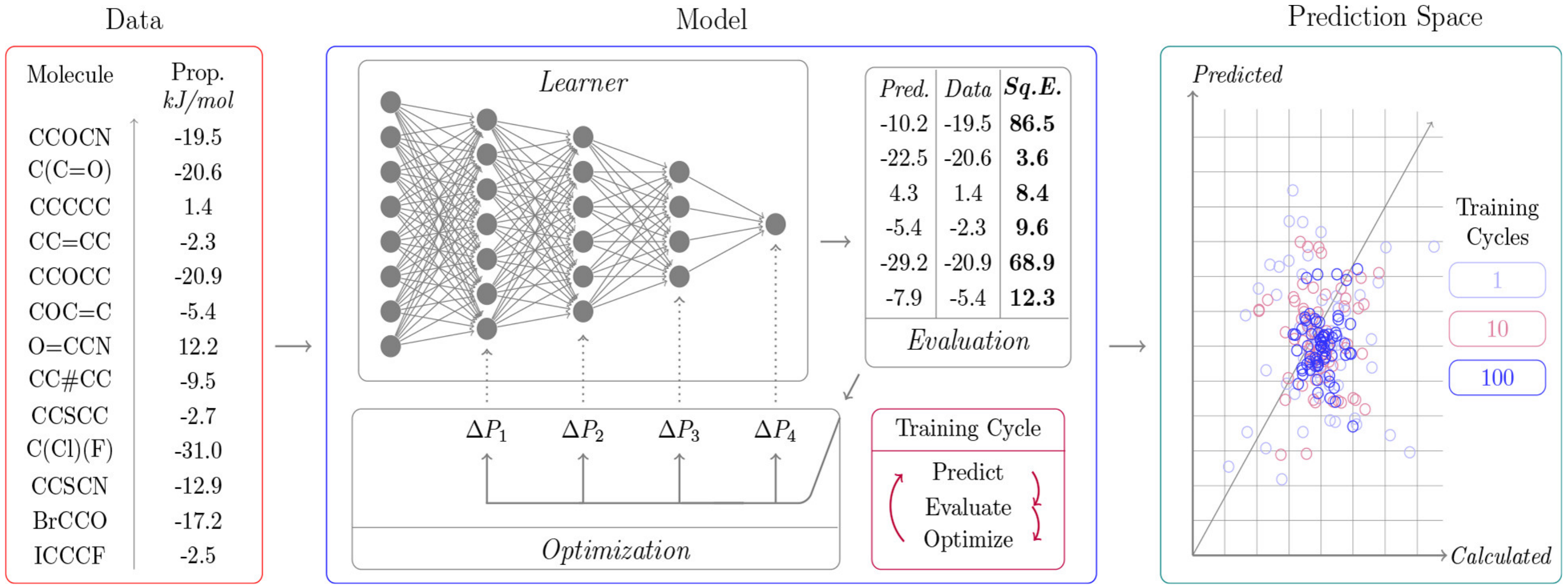
ARC Centre of Excellence for Electromaterials Science, Research School of Chemistry, Australian National University, Canberra, Australian Capital Territory 2601, Australia

ABSTRACT: Machine learning enables computers to address problems by learning from data. Deep learning is a type of machine learning that uses a hierarchical recombination of features to extract pertinent information and then learn the patterns represented in the data. Over the last eight years, its abilities have increasingly been applied to a wide variety of chemical challenges, from improving computational chemistry to drug and materials design and even synthesis planning. This review aims to explain the concepts of deep learning to chemists from any background and follows this with an overview of the diverse applications demonstrated in the literature. We hope that this will empower the broader chemical community to engage with this burgeoning field and foster the growing movement of deep learning accelerated chemistry.

KEYWORDS: Machine learning, Representation learning, Deep learning, Computational chemistry, Drug design, Materials design, Synthesis planning, Open sourcing, Quantum mechanical calculations, Cheminformatics

https://doi.org/10.1021/acs.jcim.9b00266

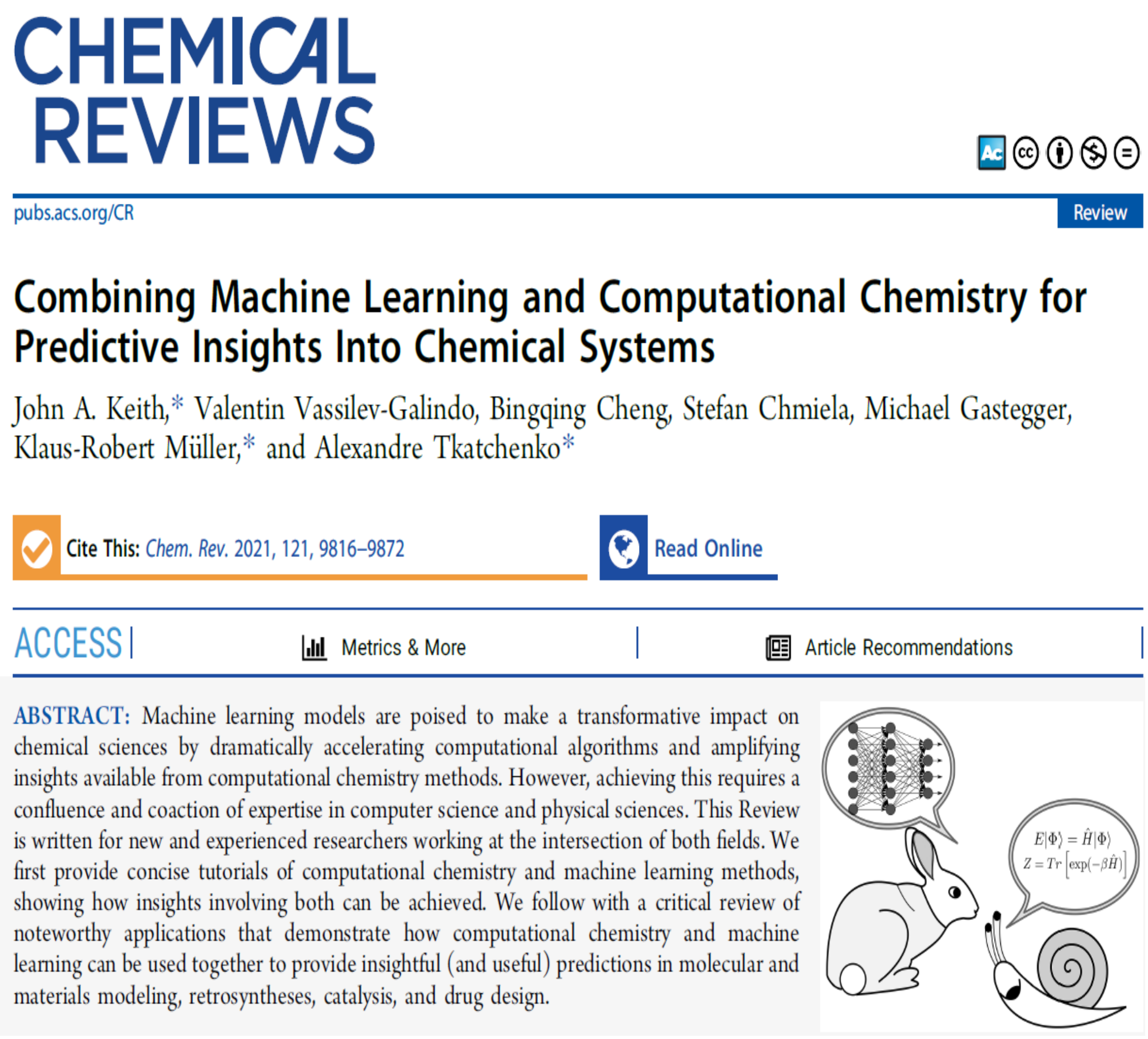
“Direct” ML-based property prediction



Molecule	Prop. *kJ/mol*
CCOCN	-19.5
C(C=O)	-20.6
CCCCC	1.4
CC=CC	-2.3
CCOCC	-20.9
COC=C	-5.4
O=CCN	12.2
CC#CC	-9.5
CCSCC	-2.7
C(Cl)(F)	-31.0
CCSCN	-12.9
BrCCO	-17.2
ICCCF	-2.5

Pred.	*Data*	***Sq.E.***
-10.2	-19.5	**86.5**
-22.5	-20.6	**3.6**
4.3	1.4	**8.4**
-5.4	-2.3	**9.6**
-29.2	-20.9	**68.9**
-7.9	-5.4	**12.3**

Machine learning in chemistry



CHEMICAL REVIEWS

pubs.acs.org/CR

Review

Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems

John A. Keith,* Valentin Vassilev-Galindo, Bingqing Cheng, Stefan Chmiela, Michael Gastegger, Klaus-Robert Müller,* and Alexandre Tkatchenko*

Cite This: *Chem. Rev.* 2021, 121, 9816–9872

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Machine learning models are poised to make a transformative impact on chemical sciences by dramatically accelerating computational algorithms and amplifying insights available from computational chemistry methods. However, achieving this requires a confluence and coaction of expertise in computer science and physical sciences. This Review is written for new and experienced researchers working at the intersection of both fields. We first provide concise tutorials of computational chemistry and machine learning methods, showing how insights involving both can be achieved. We follow with a critical review of noteworthy applications that demonstrate how computational chemistry and machine learning can be used together to provide insightful (and useful) predictions in molecular and materials modeling, retrosyntheses, catalysis, and drug design.

$E|\Phi\rangle = \hat{H}|\Phi\rangle$
$Z = Tr\left[\exp(-\beta\hat{H})\right]$

https://doi.org/10.1021/acs.chemrev.1c00107

ML-based interaction potentials

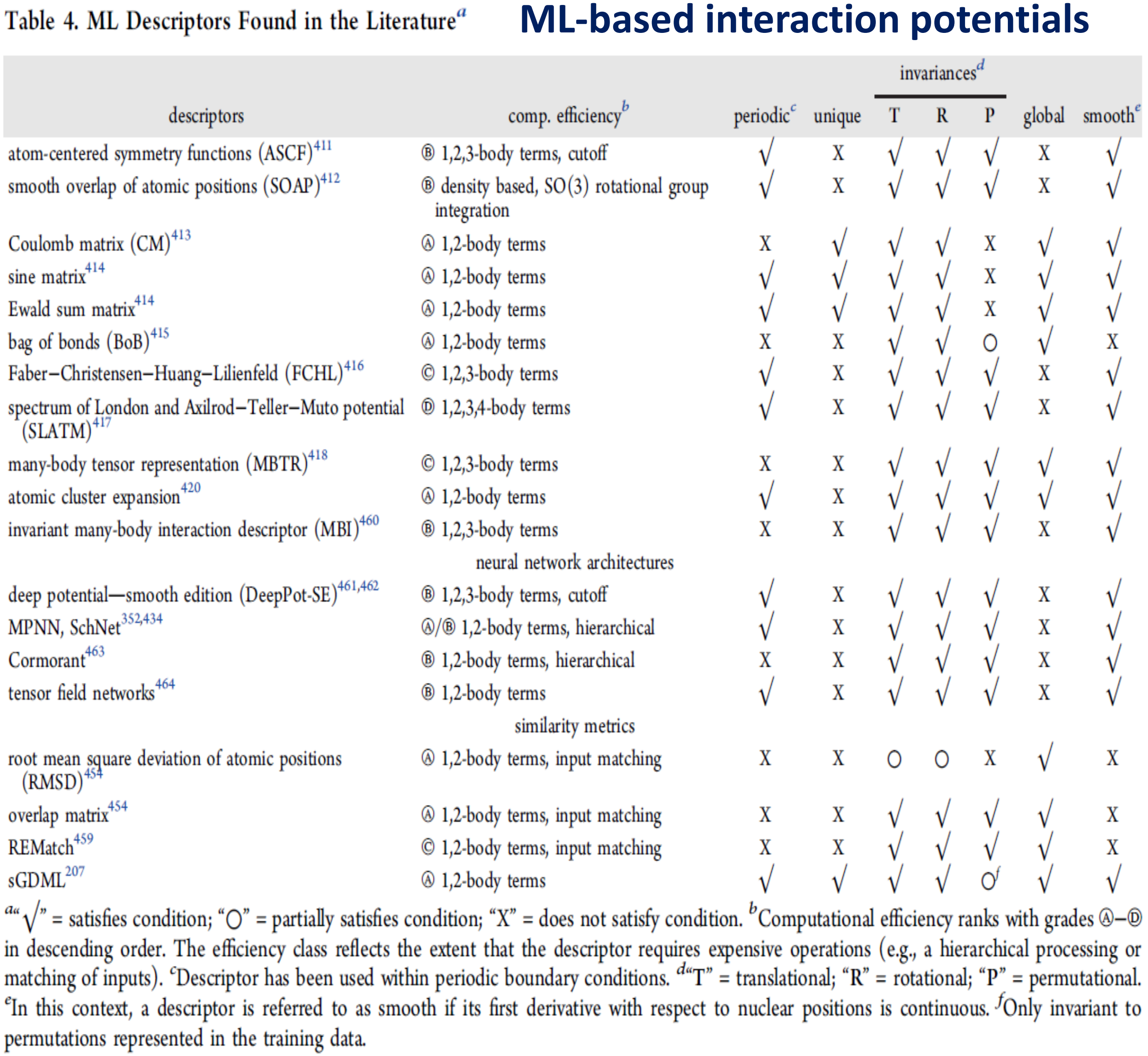
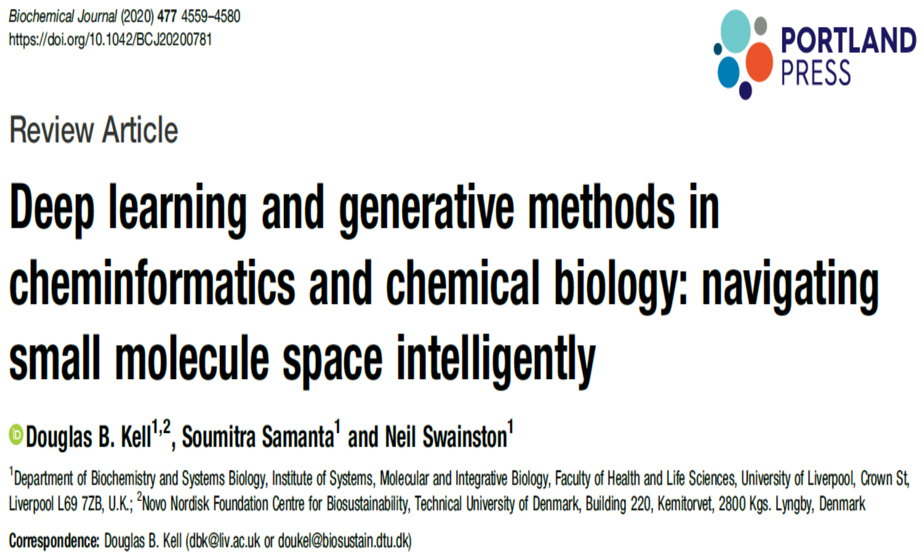


Table 4. ML Descriptors Found in the Literature[a]

descriptors	comp. efficiency[b]	periodic[c]	unique	invariances[d]			global	smooth[e]
				T	R	P		
atom-centered symmetry functions (ASCF)[411]	Ⓑ 1,2,3-body terms, cutoff	√	X	√	√	√	X	√
smooth overlap of atomic positions (SOAP)[412]	Ⓑ density based, SO(3) rotational group integration	√	X	√	√	√	X	√
Coulomb matrix (CM)[413]	Ⓐ 1,2-body terms	X	√	√	√	X	√	√
sine matrix[414]	Ⓐ 1,2-body terms	√	√	√	√	X	√	√
Ewald sum matrix[414]	Ⓐ 1,2-body terms	√	√	√	√	X	√	√
bag of bonds (BoB)[415]	Ⓐ 1,2-body terms	X	X	√	√	○	√	X
Faber–Christensen–Huang–Lilienfeld (FCHL)[416]	Ⓒ 1,2,3-body terms	√	X	√	√	√	X	√
spectrum of London and Axilrod–Teller–Muto potential (SLATM)[417]	Ⓓ 1,2,3,4-body terms	√	X	√	√	√	X	√
many-body tensor representation (MBTR)[418]	Ⓒ 1,2,3-body terms	X	X	√	√	√	√	√
atomic cluster expansion[420]	Ⓐ 1,2-body terms	√	X	√	√	√	√	√
invariant many-body interaction descriptor (MBI)[460]	Ⓑ 1,2,3-body terms	X	X	√	√	√	X	√
	neural network architectures							
deep potential—smooth edition (DeepPot-SE)[461,462]	Ⓑ 1,2,3-body terms, cutoff	√	X	√	√	√	X	√
MPNN, SchNet[352,434]	Ⓐ/Ⓑ 1,2-body terms, hierarchical	√	X	√	√	√	X	√
Cormorant[463]	Ⓑ 1,2-body terms, hierarchical	X	X	√	√	√	X	√
tensor field networks[464]	Ⓑ 1,2-body terms	√	X	√	√	√	X	√
	similarity metrics							
root mean square deviation of atomic positions (RMSD)[454]	Ⓐ 1,2-body terms, input matching	X	X	○	○	X	√	X
overlap matrix[454]	Ⓐ 1,2-body terms, input matching	X	X	√	√	√	√	X
REMatch[459]	Ⓒ 1,2-body terms, input matching	X	X	√	√	√	√	X
sGDML[207]	Ⓐ 1,2-body terms	√	√	√	√	○[f]	√	√

[a] "√" = satisfies condition; "○" = partially satisfies condition; "X" = does not satisfy condition. [b] Computational efficiency ranks with grades Ⓐ–Ⓓ in descending order. The efficiency class reflects the extent that the descriptor requires expensive operations (e.g., a hierarchical processing or matching of inputs). [c] Descriptor has been used within periodic boundary conditions. [d] "T" = translational; "R" = rotational; "P" = permutational. [e] In this context, a descriptor is referred to as smooth if its first derivative with respect to nuclear positions is continuous. [f] Only invariant to permutations represented in the training data.

Machine learning in chemistry



Biochemical Journal (2020) **477** 4559–4580
https://doi.org/10.1042/BCJ20200781

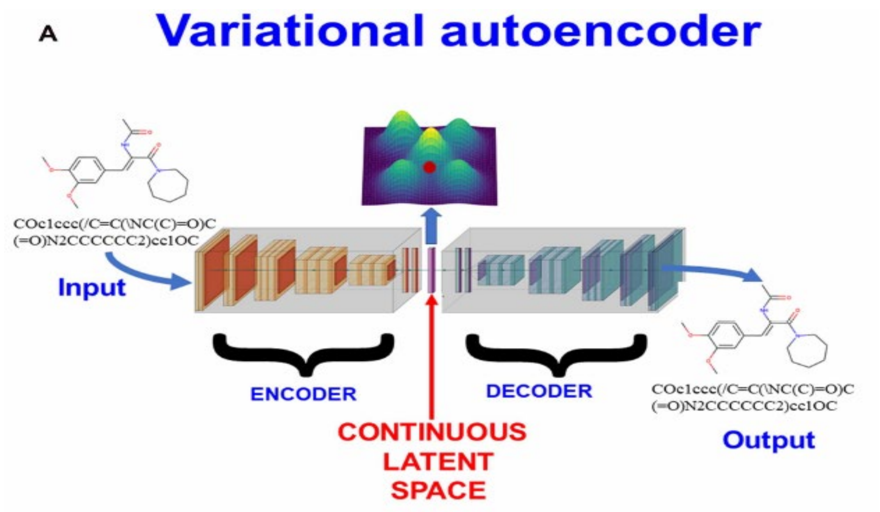
Review Article

Deep learning and generative methods in cheminformatics and chemical biology: navigating small molecule space intelligently

Douglas B. Kell[1,2], Soumitra Samanta[1] and Neil Swainston[1]

[1]Department of Biochemistry and Systems Biology, Institute of Systems, Molecular and Integrative Biology, Faculty of Health and Life Sciences, University of Liverpool, Crown St, Liverpool L69 7ZB, U.K.; [2]Novo Nordisk Foundation Centre for Biosustainability, Technical University of Denmark, Building 220, Kemitorvet, 2800 Kgs. Lyngby, Denmark

Correspondence: Douglas B. Kell (dbk@liv.ac.uk or doukel@biosustain.dtu.dk)


A
Variational autoencoder
COc1ccc(/C=C(\NC(C)=O)C(=O)N2CCCCCC2)cc1OC
Input
ENCODER
DECODER
CONTINUOUS LATENT SPACE
COc1ccc(/C=C(\NC(C)=O)C(=O)N2CCCCCC2)cc1OC
Output

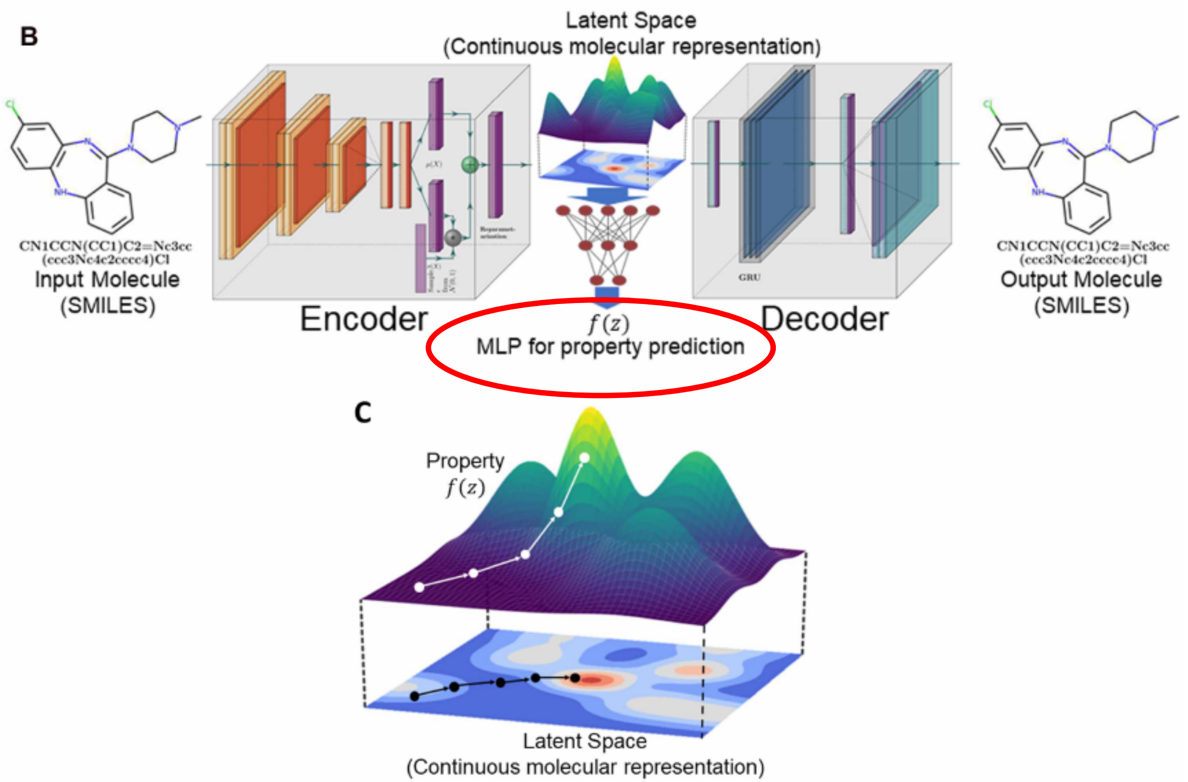

B
Latent Space
(Continuous molecular representation)
CN1CCN(CC1)C2=Nc3cc(ccc3Nc4c2cccc4)Cl
Input Molecule
(SMILES)
Encoder
f(z)
MLP for property prediction
Decoder
GRU
CN1CCN(CC1)C2=Nc3cc(ccc3Nc4c2cccc4)Cl
Output Molecule
(SMILES)
C
Property
f(z)
Latent Space
(Continuous molecular representation)

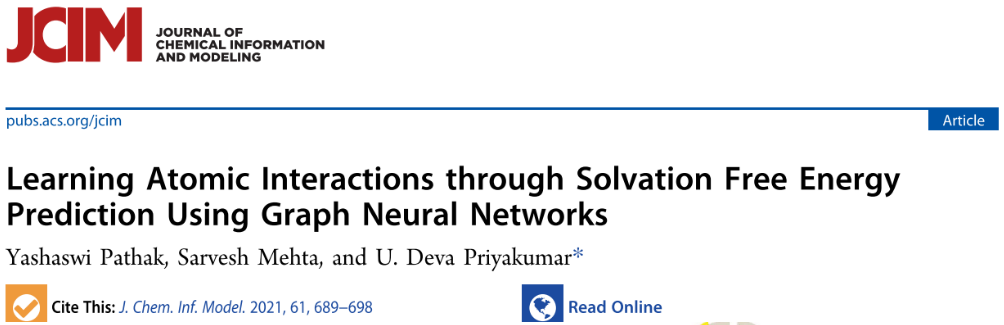
Figure 5. Variational autoencoder networks and their uses.
(**A**) Basic VAE architecture, showing the latent space. (**B**) VAE as proposed by Gómez-Bombarelli and colleagues [27]. The latent space is shown as a 2D space for ease of visualisation, but in the paper had a dimensionality of either 156 or (more commonly) 196. (**C**) Moving around the latent space, one simultaneously comes into the 'basin of attraction' of particular molecules, whose structures may be output and properties may be calculated via the MLP shown in (**A**) and described in the text (based on [27]). Using optimisation strategies such as evolutionary algorithms can guide the search for the properties and hence the 'novel' molecules.

https://doi.org/10.1042/BCJ20200781

Machine learning in chemistry

ML in solvation science

- An underexplored field
- Focused on solvation free energies (pure ML or physics-augmented ML regression)
- Equation-of-state modeling
- Data interpolation / extension



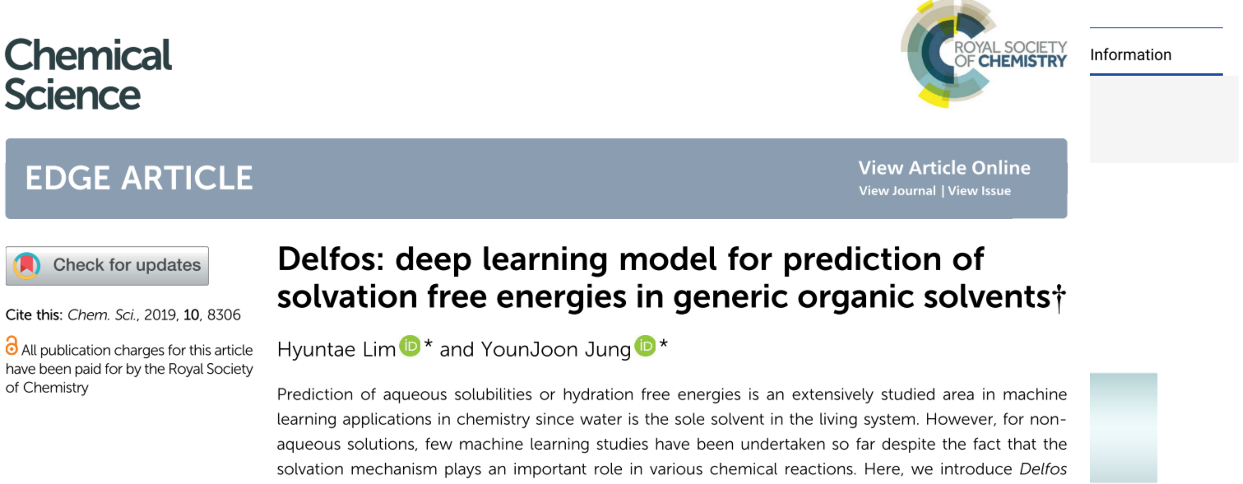
JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

Learning Atomic Interactions through Solvation Free Energy Prediction Using Graph Neural Networks

Yashaswi Pathak, Sarvesh Mehta, and U. Deva Priyakumar*

Cite This: *J. Chem. Inf. Model.* 2021, 61, 689–698 Read Online



Chemical Science ROYAL SOCIETY OF CHEMISTRY Information

EDGE ARTICLE View Article Online View Journal | View Issue

Check for updates

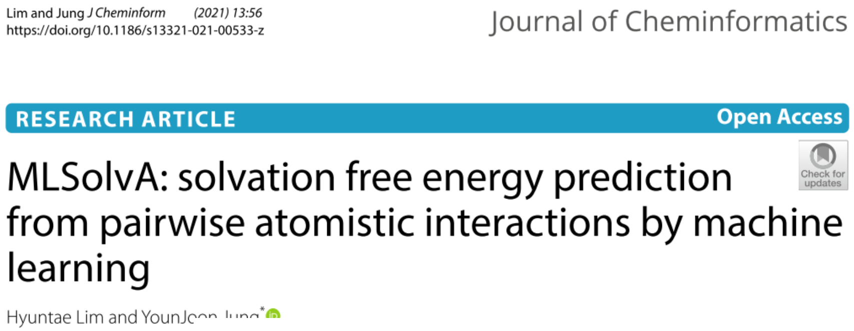
Cite this: *Chem. Sci.*, 2019, **10**, 8306

All publication charges for this article have been paid for by the Royal Society of Chemistry

Delfos: deep learning model for prediction of solvation free energies in generic organic solvents†

Hyuntae Lim * and YounJoon Jung *

Prediction of aqueous solubilities or hydration free energies is an extensively studied area in machine learning applications in chemistry since water is the sole solvent in the living system. However, for non-aqueous solutions, few machine learning studies have been undertaken so far despite the fact that the solvation mechanism plays an important role in various chemical reactions. Here, we introduce *Delfos*



Lim and Jung *J Cheminform* (2021) 13:56
https://doi.org/10.1186/s13321-021-00533-z

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

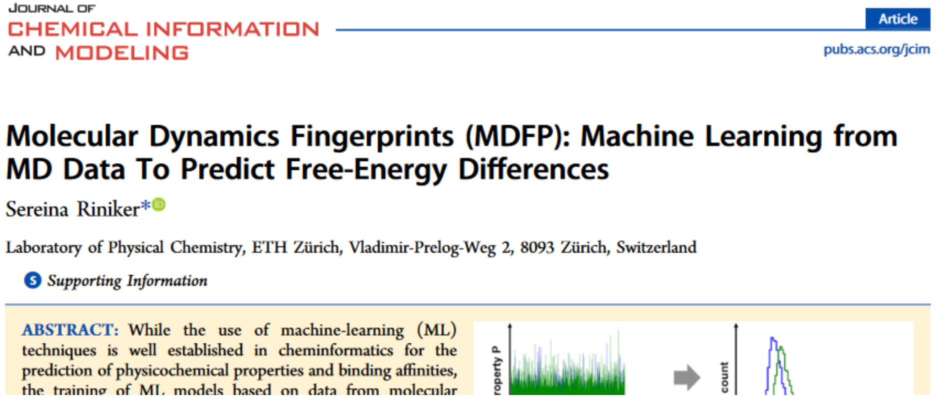
Check for updates

MLSolvA: solvation free energy prediction from pairwise atomistic interactions by machine learning

Hyuntae Lim and YounJoon Jung*

Abstract

Recent advances in mach
structure–property relatic
we introduce a novel ML-



JOURNAL OF CHEMICAL INFORMATION AND MODELING Article

pubs.acs.org/jcim

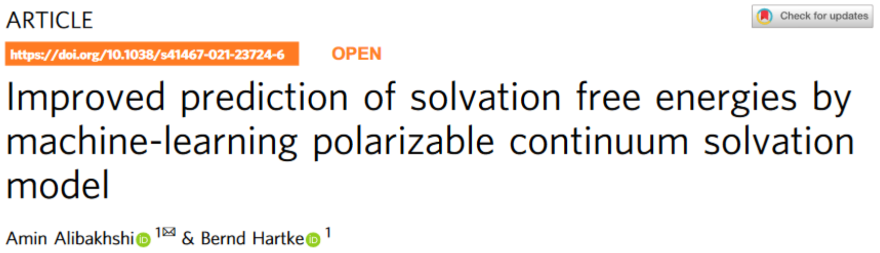
Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data To Predict Free-Energy Differences

Sereina Riniker*

Laboratory of Physical Chemistry, ETH Zürich, Vladimir-Prelog-Weg 2, 8093 Zürich, Switzerland

Supporting Information

ABSTRACT: While the use of machine-learning (ML) techniques is well established in cheminformatics for the prediction of physicochemical properties and binding affinities, the training of ML models based on data from molecular



ARTICLE Check for updates

https://doi.org/10.1038/s41467-021-23724-6 OPEN

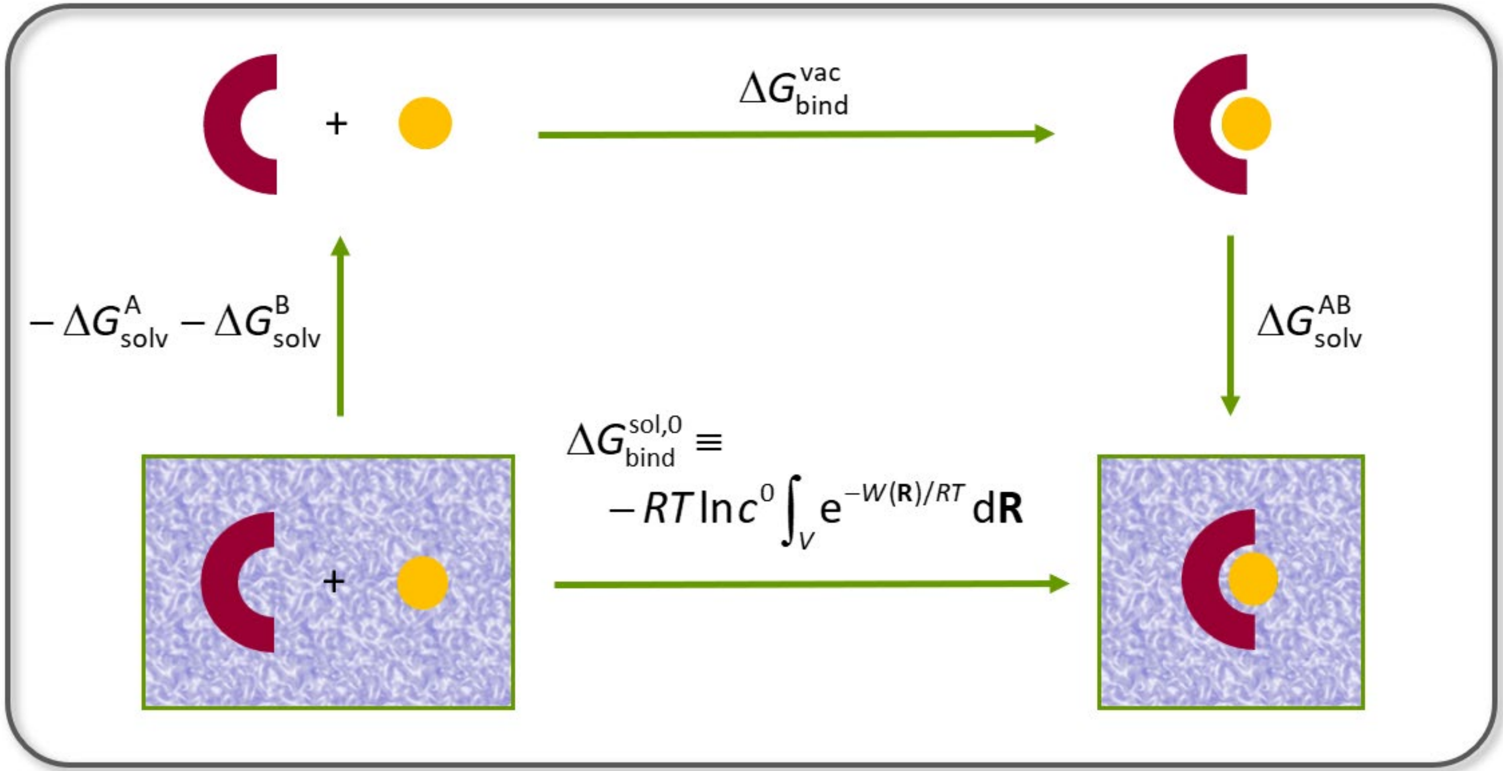
Improved prediction of solvation free energies by machine-learning polarizable continuum solvation model

Amin Alibakhshi [1]✉ & Bernd Hartke [1]

Thermodynamic properties in chemistry

Free Energies are key to processes in solution

- Binding
- Folding
- Reactions
- …



$$\Delta G^{\mathrm{vac}}_{\mathrm{bind}}$$

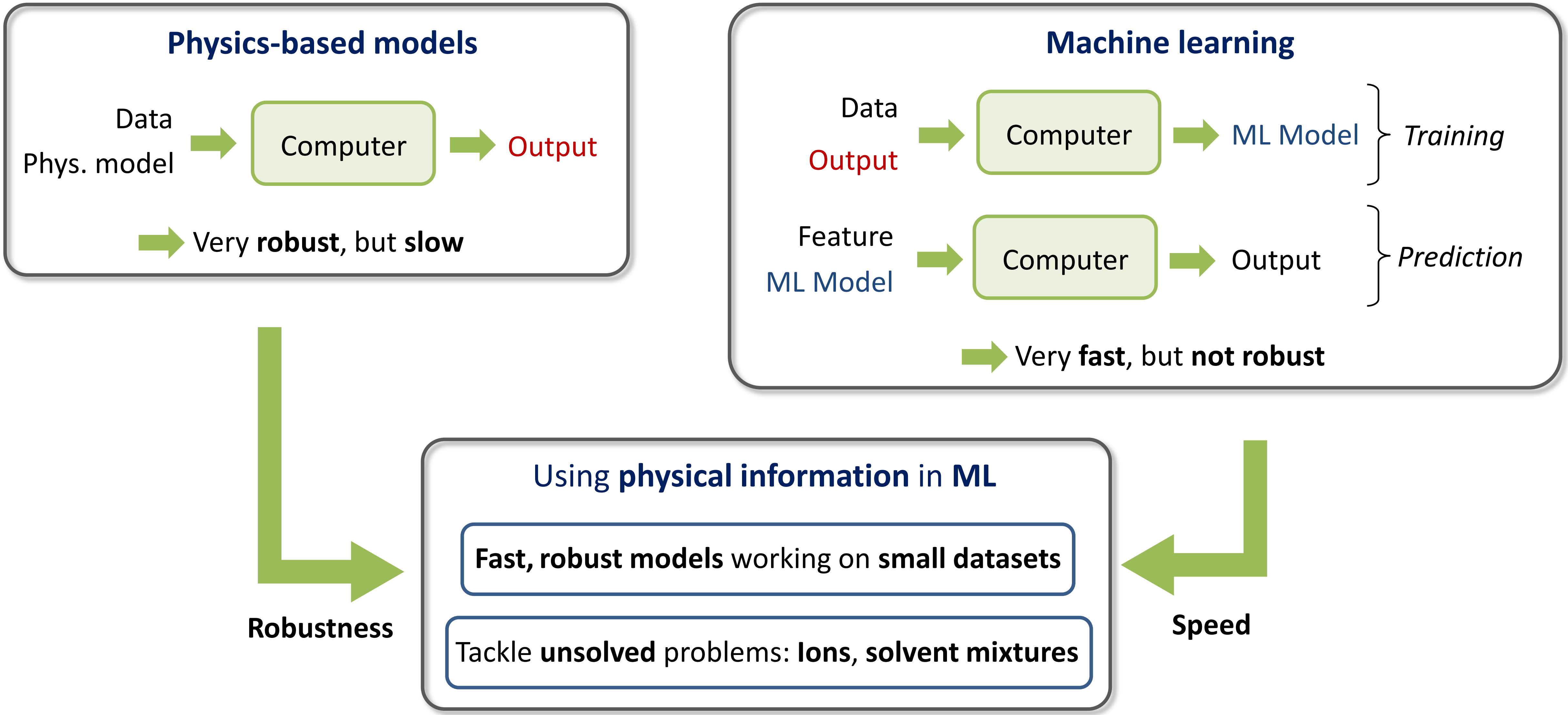
$$-\Delta G^{\mathrm{A}}_{\mathrm{solv}} - \Delta G^{\mathrm{B}}_{\mathrm{solv}}$$

$$\Delta G^{\mathrm{AB}}_{\mathrm{solv}}$$

$$\Delta G^{\mathrm{sol,0}}_{\mathrm{bind}} \equiv -RT\ln c^{0}\int_{V} \mathrm{e}^{-W(\mathbf{R})/RT}\,\mathrm{d}\mathbf{R}$$

$$\boxed{\Delta G^{\mathrm{sol}}_{\mathrm{bind}} = -\Delta G^{\mathrm{A}}_{\mathrm{solv}} - \Delta G^{\mathrm{B}}_{\mathrm{solv}} + \Delta G^{\mathrm{vac}}_{\mathrm{bind}} + \Delta G^{\mathrm{AB}}_{\mathrm{solv}}}$$

The best of both worlds?



Physics-based models

Data

Phys. model → Computer → Output

→ Very **robust**, but **slow**

Machine learning

Data

Output → Computer → ML Model } *Training*

Feature

ML Model → Computer → Output } *Prediction*

→ Very **fast**, but **not robust**

Robustness

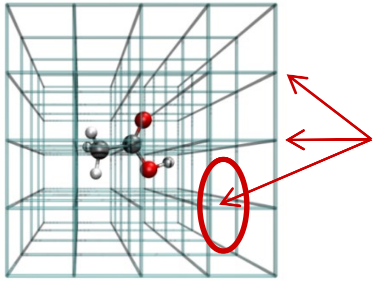
Speed

Using **physical information** in **ML**

Fast, robust models working on **small datasets**

Tackle **unsolved** problems: **Ions**, **solvent mixtures**

Thermodynamic basis: 3D reference interaction site model (3DRISM)



Basic idea

- Solute-molecule/solvent-atom (γ) interaction on grid
- Coupled nonlinear integral equation/closure

Distribution functions

$$g(\mathbf{r}) = h(\mathbf{r}) + 1$$

$$\boldsymbol{\chi} = \boldsymbol{\rho\omega} + \boldsymbol{\rho h \rho}$$

$$\mu^{\text{ex}} = -\beta^{-1}\rho_0 \sum_{\alpha}\sum_{\gamma}\int_0^1 d\lambda \int d\mathbf{r}\; g_\gamma(\mathbf{r},\lambda)\frac{\partial u_{\alpha\gamma}(\mathbf{r},\lambda)}{\partial\lambda}$$

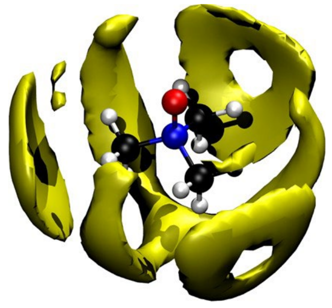
$$\rho_\gamma h_\gamma(\mathbf{r}) = \sum_{\gamma'}\int d\mathbf{r}' c_\gamma(\mathbf{r}')\, \chi_{\gamma'\gamma}(\mathbf{r}-\mathbf{r}')$$

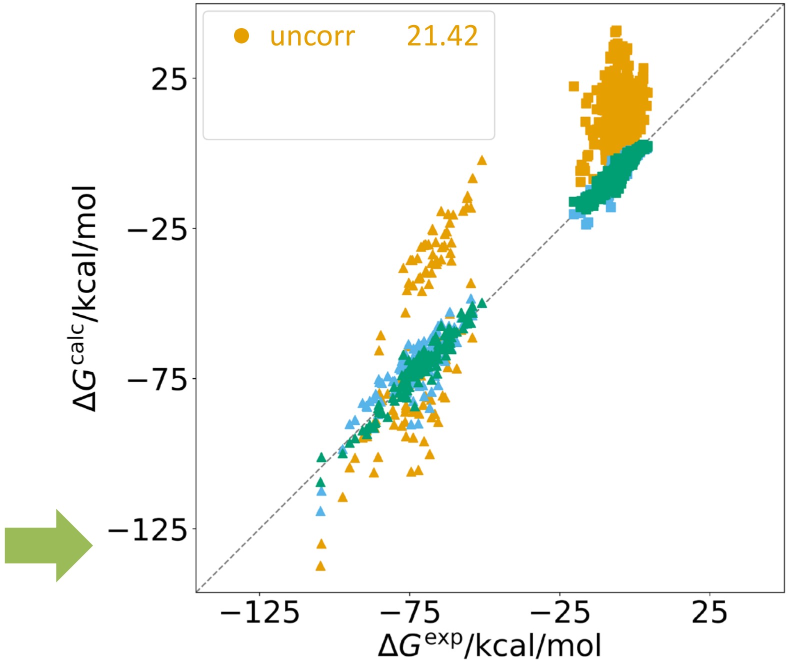
Partial molar volume

$$V^{\text{m}} = F[c]$$

Local free energies (LFE)

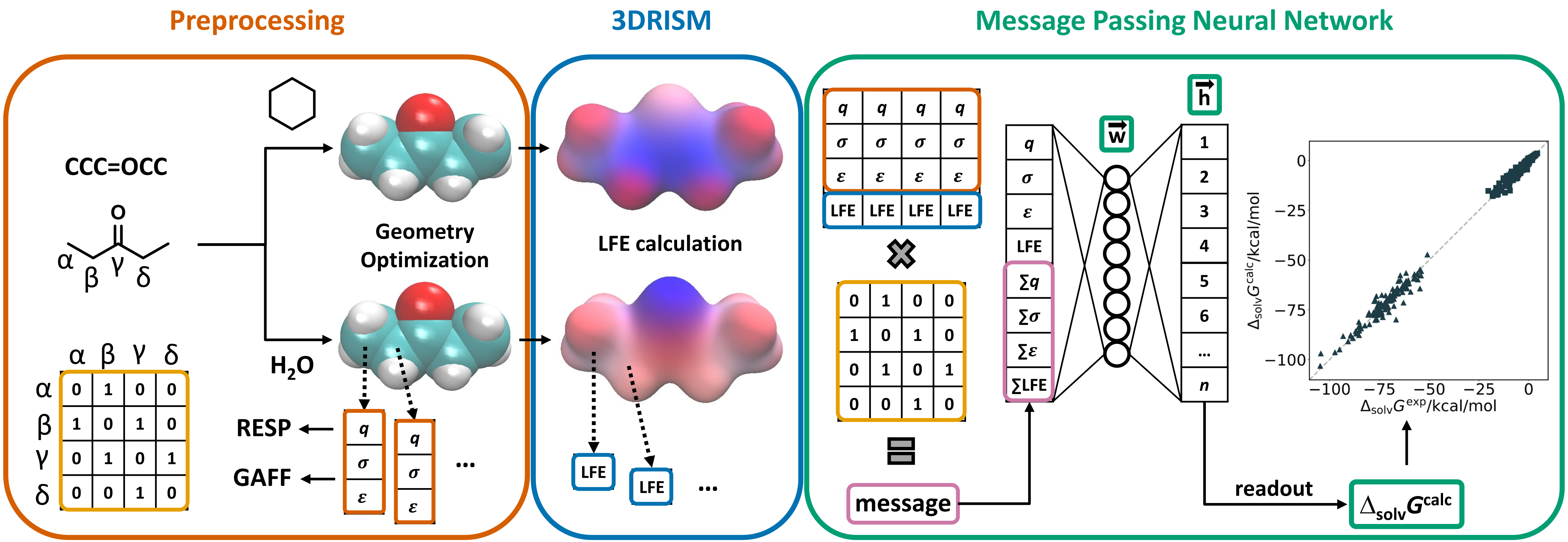
EC




● uncorr 21.42
25
−25
−75
−125
ΔG^{calc}/kcal/mol
−125
−75
−25
25
ΔG^{exp}/kcal/mol

S. M. Kast, T. Kloss, *J. Chem. Phys.* **2008**, *129*.

ML approach: Message passing neural network (MPNN)



	α	β	γ	δ
α	0	1	0	0
β	1	0	1	0
γ	0	1	0	1
δ	0	0	1	0

Adjacency matrix

Lennard-Jones-parameters

Local free energies

Feature matrix

Message matrix

Neural net

Predicted solvation free energy

J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, *34th Int. Conf. Mach. Learn. ICML 2017* **2017**, *3*, 2053–2070.

Method comparison – Is the ML model usable?

🟧 ML: Proposed method

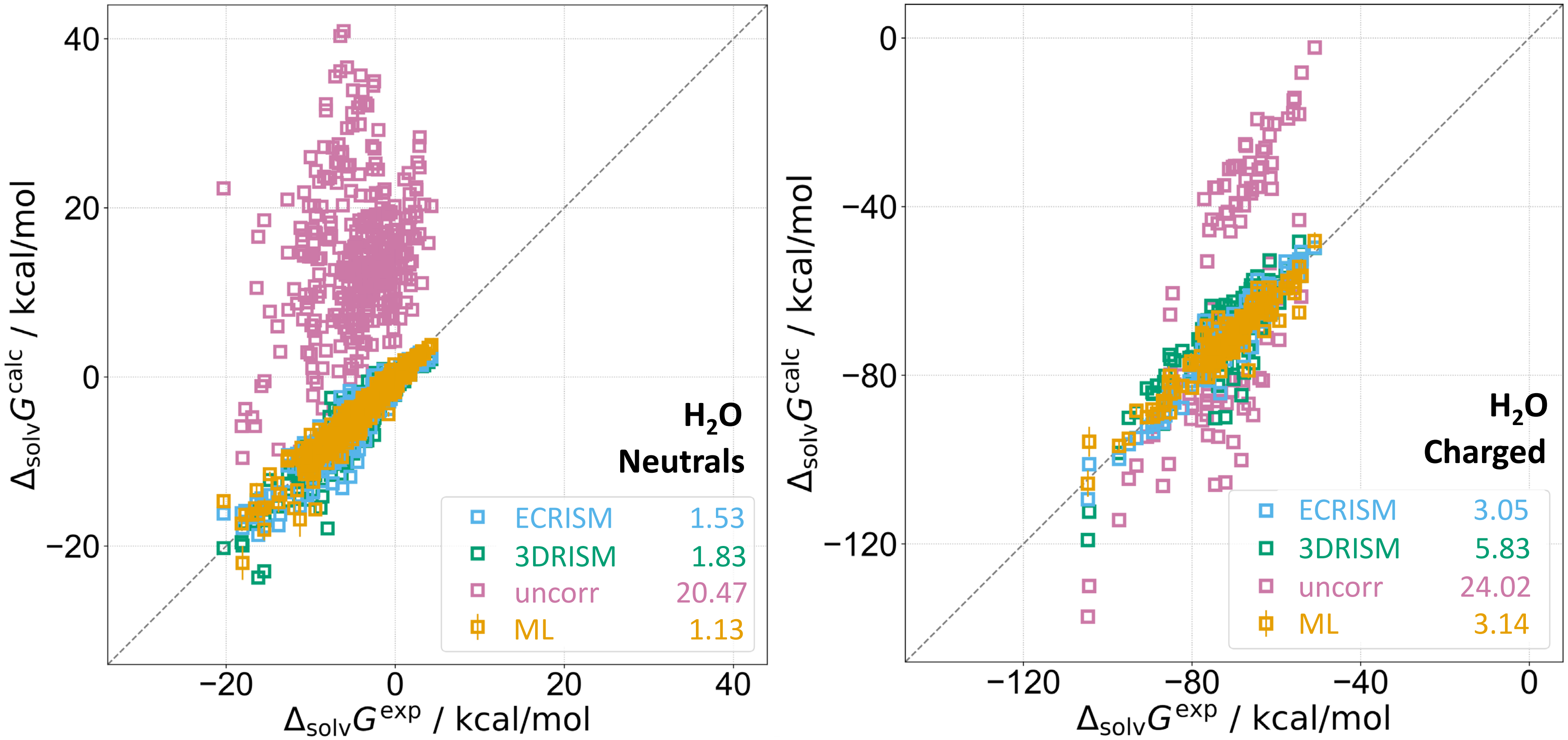
🟦 ECRISM: State-of-the-art quantum chemical model

🟩 3DRISM: Corrected reference method

🟪 Uncorr: Thermodynamic basis for the ML model

- Small RMSE for 🟧 overall, especially for neutrals
- Comparable results between 🟧 and 🟦 for charged subset

5-fold **crossvalidated** results



N. Tielker, D. Tomazic, J. Heil, T. Kloss, S. Ehrhart, S. Güssregen, *J. Comput.-Aided Mol. Des.* 30, 1035 (2016)
N. Tielker, L. Eberlein, S. Güssregen, S. M. Kast, *J. Comput.-Aided Mol. Des.* 32, 1151 (2018)

LFE info value comparison

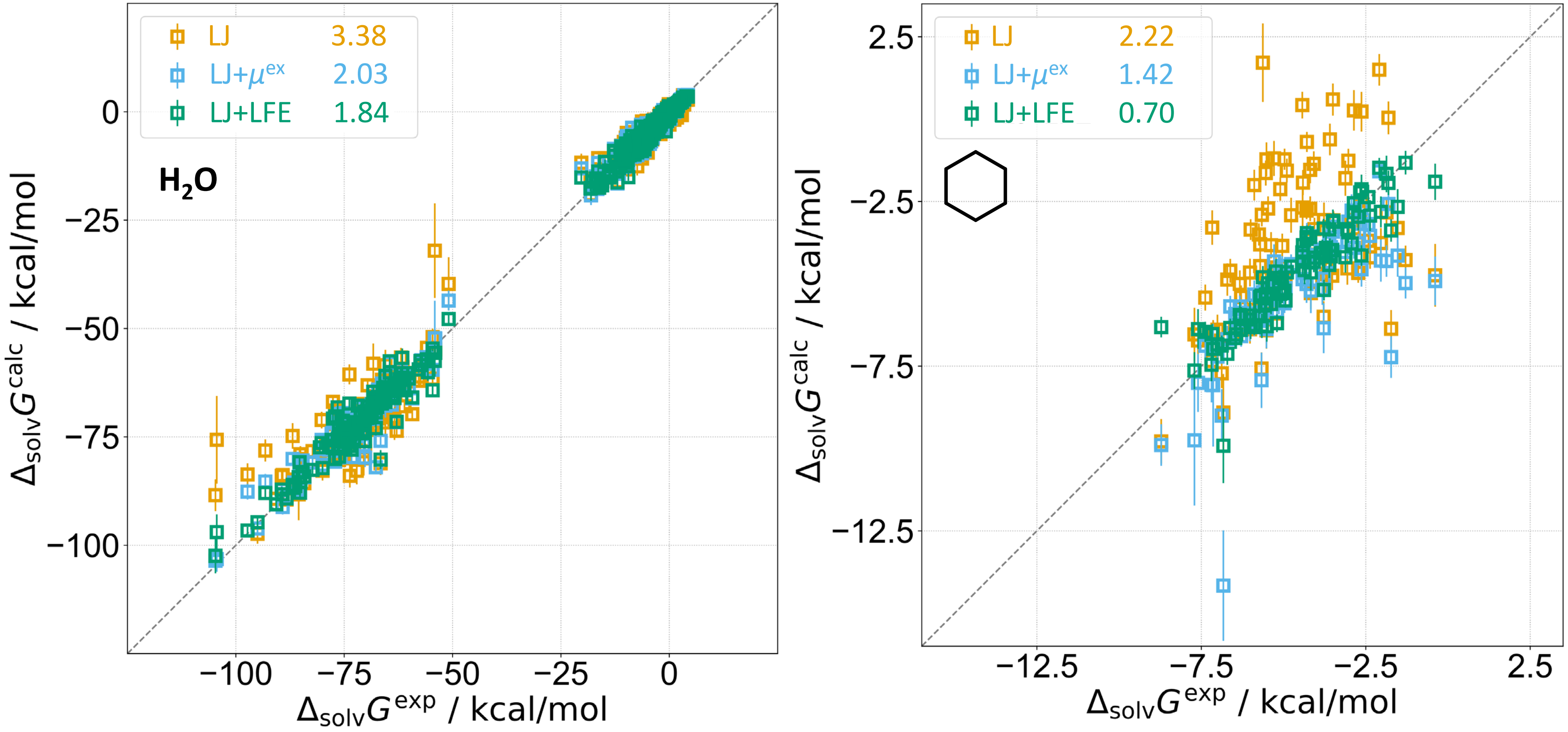
🟧 LJ: Zero-hypothesis model

🟦 LJ+μ^{ex}: Thermodynamic info **not local**

🟩 LJ+LFE: **Local** thermodynamic info

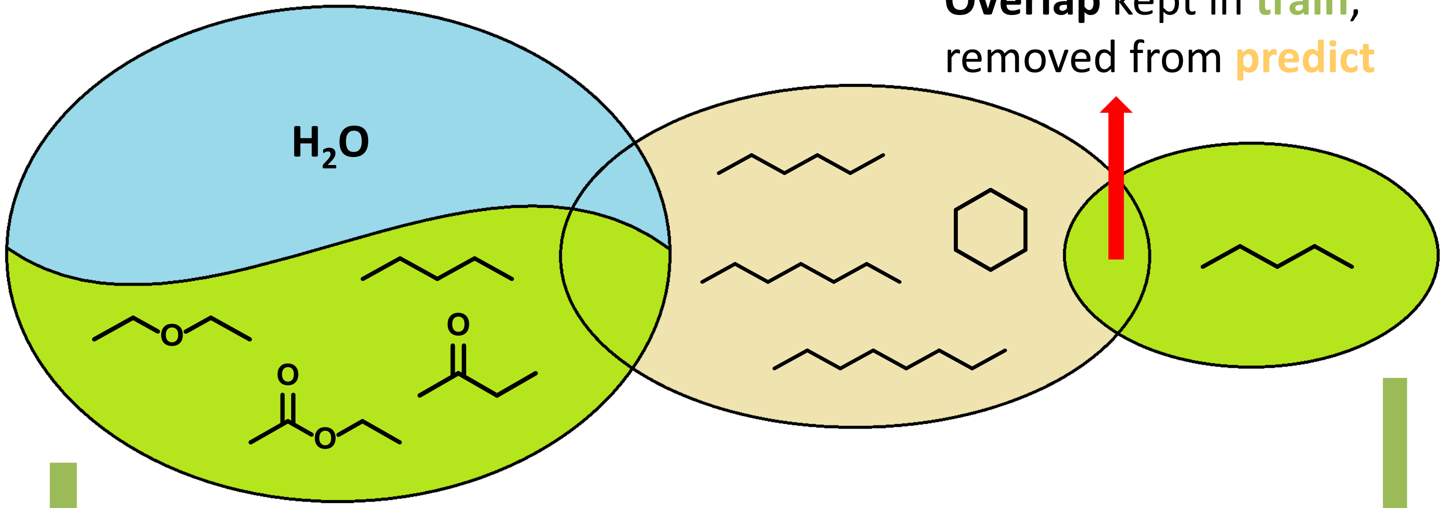
- As expected, 🟧 performs the worst
- Comparison of 🟦 and 🟩 shows **localization advantage**, especially prominent in ⬡

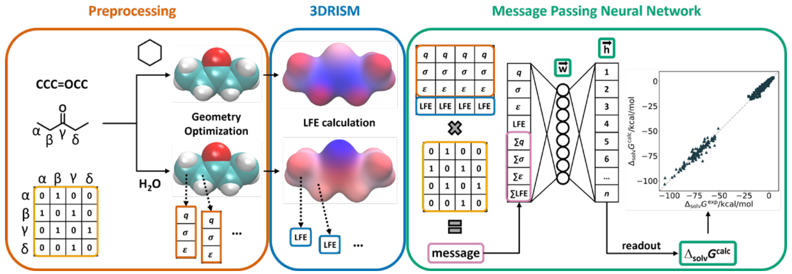
Same model trained on H_2O and ⬡ data **simultaneously**, 5-fold cv results



Transfer learning – blind prediction test

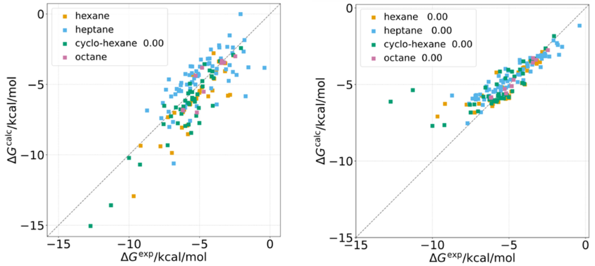
Overlap kept in train,
removed from predict

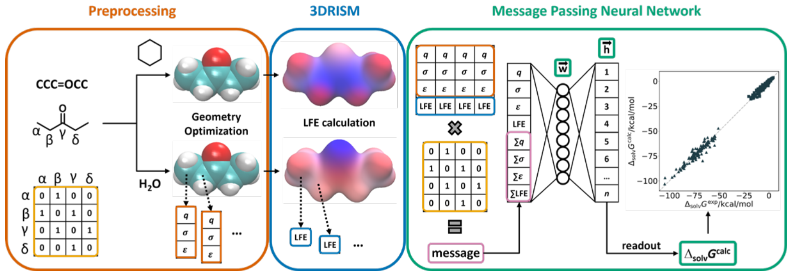




Model

Blind predictions



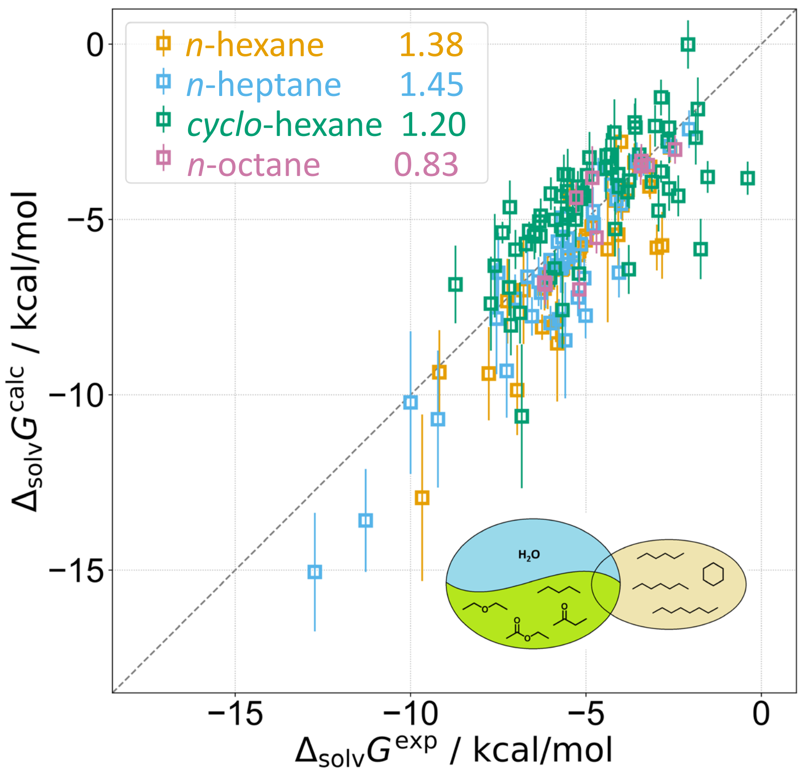


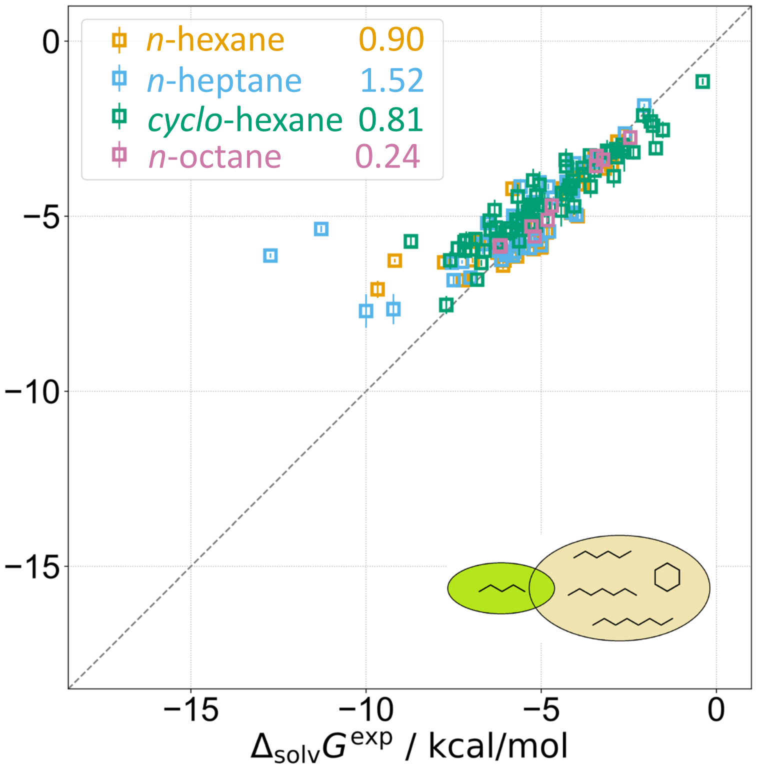
Model

Transfer learning – blind prediction test

Overall good RMSE of **1.33** kcal/mol

- Distribution more spread out, especially for 🟩
- No clear outliers in any solvent

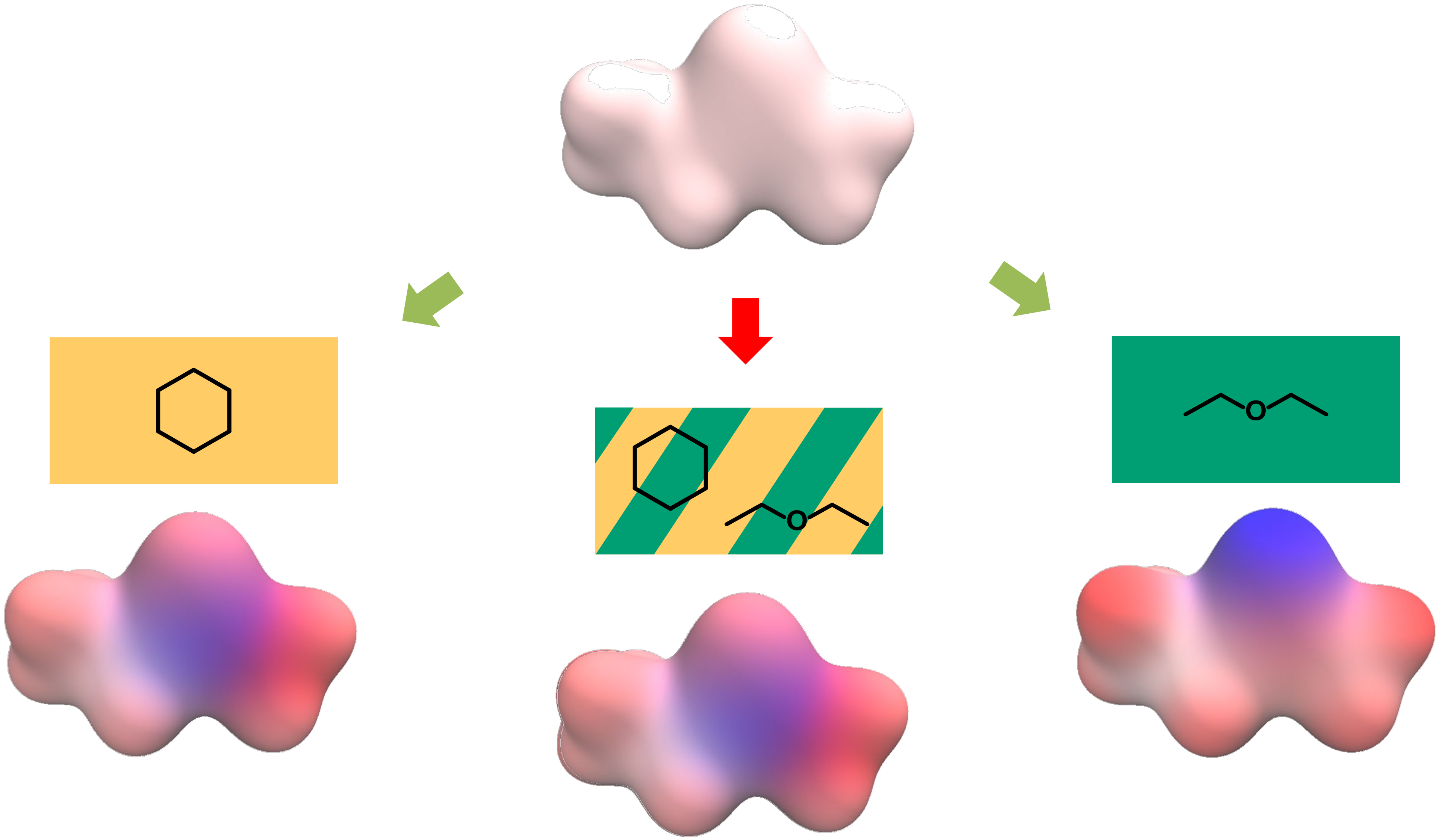




Overall excellent RMSE of **1.05** kcal/mol

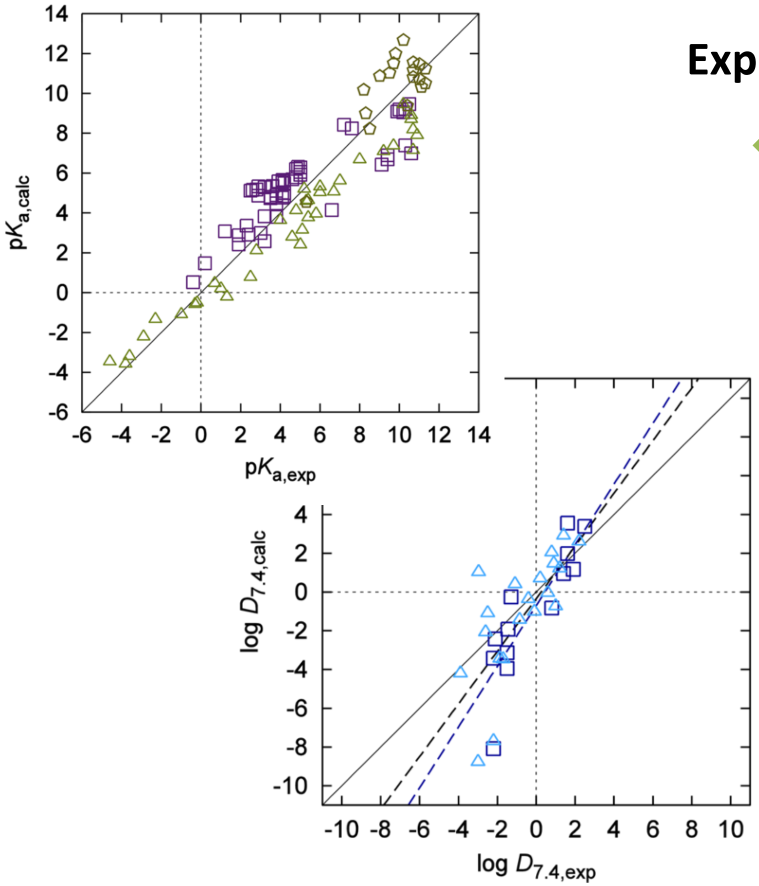
- Superb agreement with experiment for 🟪
- Some outliers for 🟦 and 🟧

Outlook – solvent mixtures

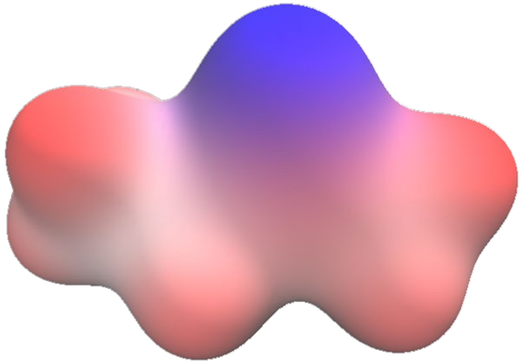


Outlook – LFE possibilities

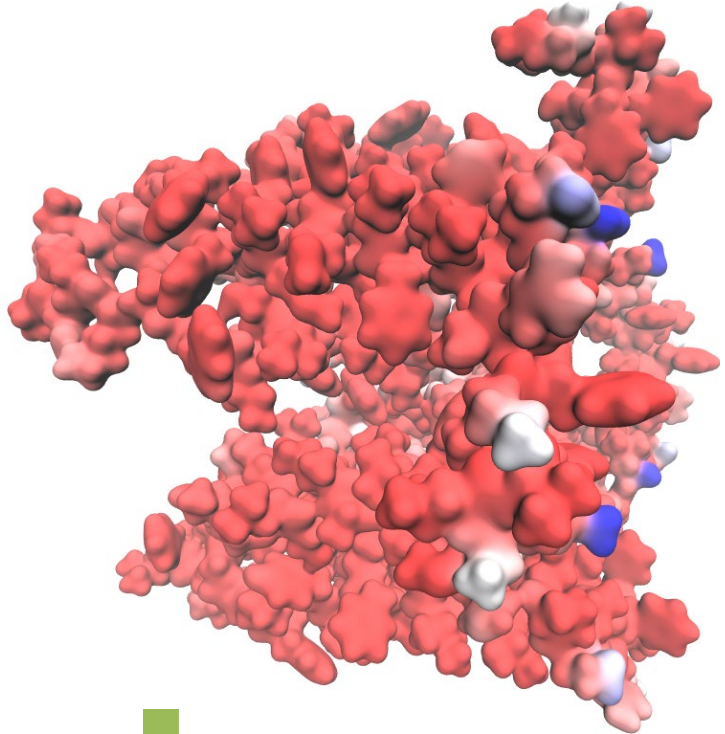
pK_a, log$D$, ...



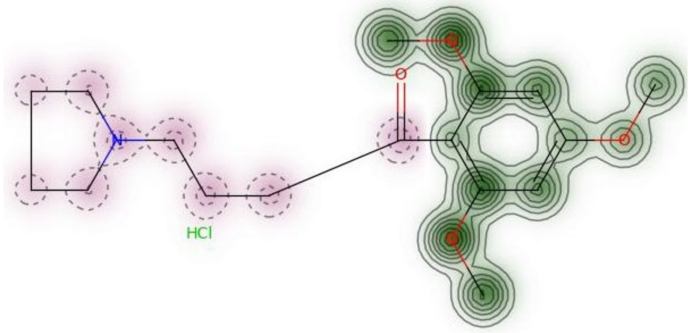
Expanding

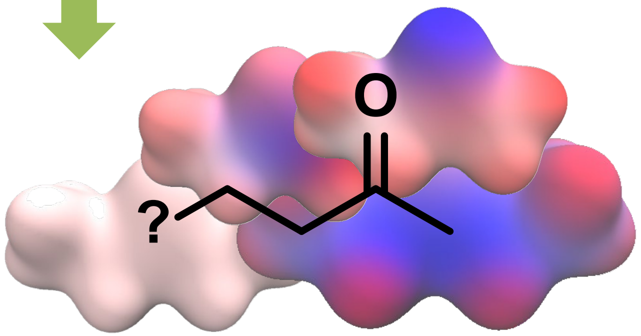


Scoring



Matching





Generative Chemistry

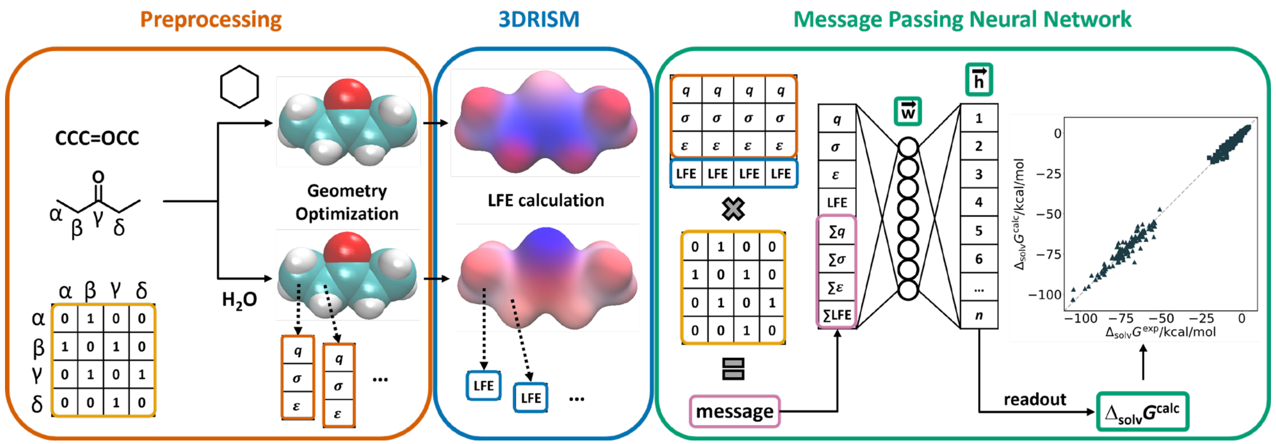
N. Tielker *et al.*, *J. Comput. Aided. Mol. Des.* **2016**, *30*, 1035–1044.
https://www.programmersought.com/article/49485546708/

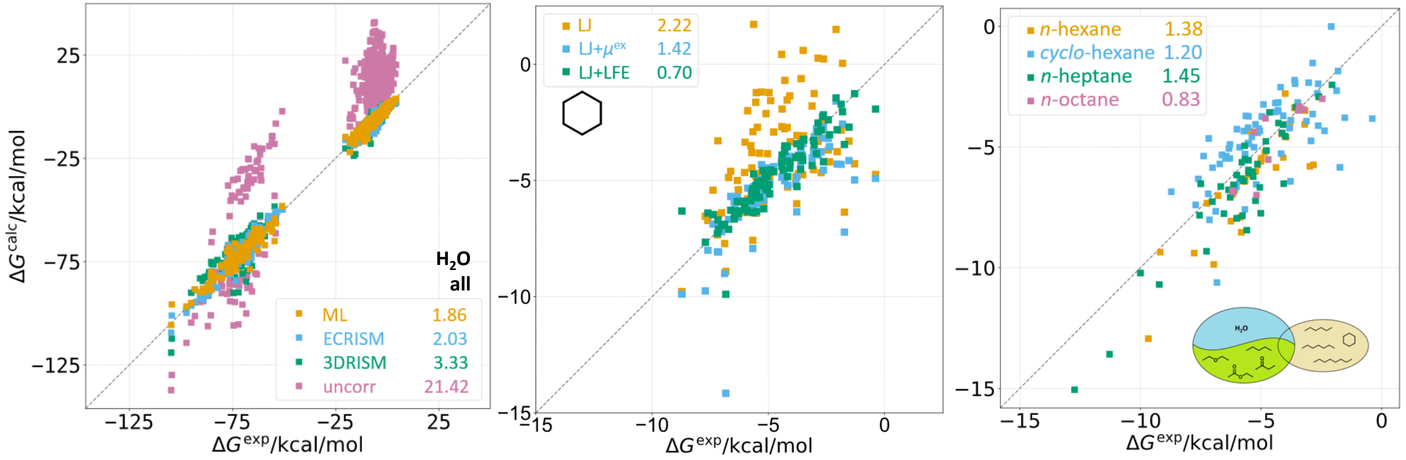
Conclusion and outlook

- **Physics based features** used in MPNN
  - ➢ Lennard-Jones-Parameters
  - ➢ Local free energies
- **Small RMSE** for whole MNSOL dataset in water
- **RMSE comparable** to **QM** for **charged** subset
- **LFE** are **valuable feature** for ML
- **Blind predictions** perform **well**
  - ➢ Applicable to **unknown solvents**

Next step: **solvent mixtures**

PC-SAFT, current **RESOLV cooperation**



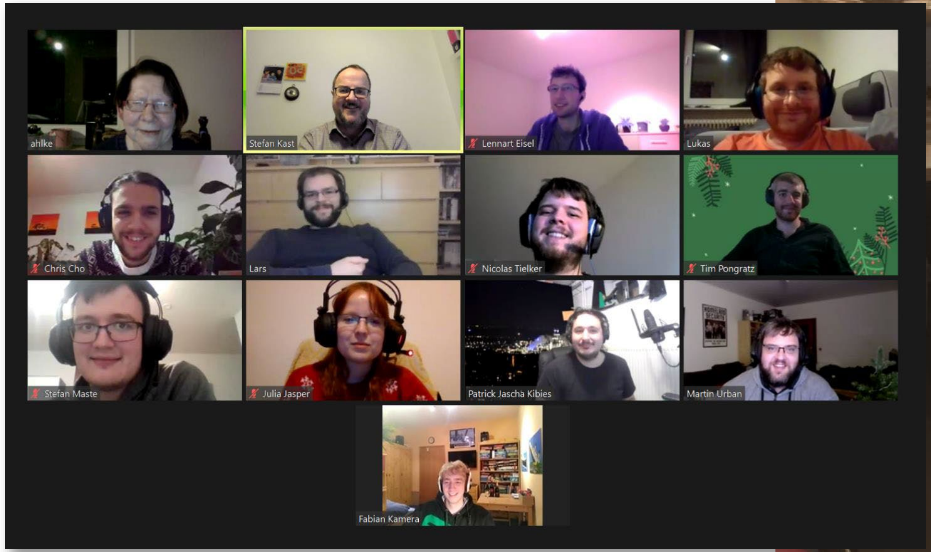


Thanks!

Dr. Yannic Alber

Dr. Julia Jasper

Kast work group



technische universität dortmund

fakultät für chemie und chemische biologie

RESOLV RUHR EXPLORES SOLVATION CLUSTER OF EXCELLENCE - EXC 2033

Localized free energies (LFE)

Chemical excess potential

$$\mu^{ex} = \beta^{-1}\rho\int d\mathbf{r}\left(h^2/2 - c\; - hc/2\right) + F[B,\bar{V}]$$

Distribution functions

$$g(\mathbf{r}) = h(\mathbf{r}) + 1$$

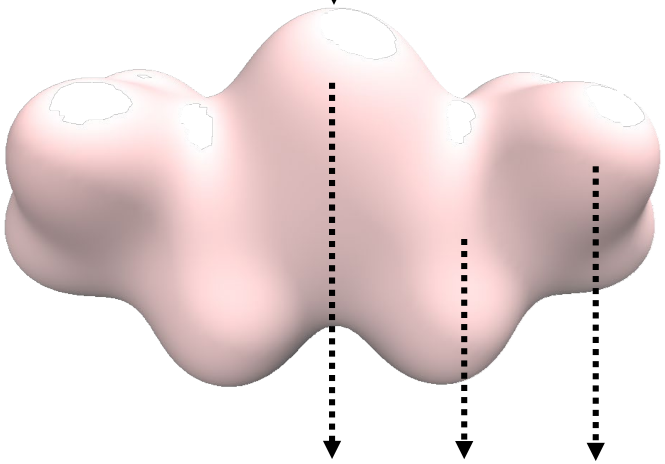
3DRISM formalism

$$h_\gamma(\mathbf{r}) = \exp[-\beta\, u_\gamma(\mathbf{r}) + h_\gamma(\mathbf{r}) - c_\gamma(\mathbf{r}) + B_\gamma(\mathbf{r})] - 1$$

$$-\beta\mu^{ex} = \rho_0 \sum_\alpha \sum_\gamma \int_0^1 d\lambda \int d\mathbf{r}\, g_{\alpha\gamma}(\mathbf{r},\lambda)\frac{\partial u_{\alpha\gamma}(\mathbf{r},\lambda)}{\partial\lambda}$$

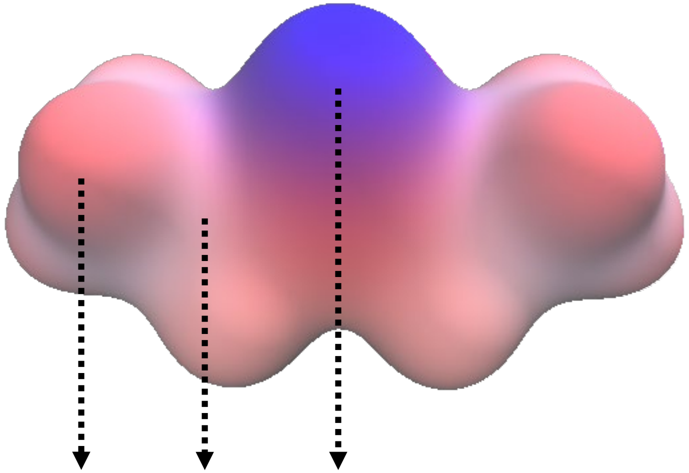
$$\mu_\alpha^{ex} = -\beta^{-1}\rho_0 \sum_\gamma \int_0^1 d\lambda \int d\mathbf{r}\, g_{\alpha\gamma}(\mathbf{r},\lambda)\frac{\partial u_{\alpha\gamma}(\mathbf{r},\lambda)}{\partial\lambda}$$

$$\sum^n \frac{\mu^{ex}}{n} = \mu^{ex} = \sum_\alpha^n \mu_\alpha^{ex}$$



$$\frac{\mu^{ex}}{n} = \frac{\mu^{ex}}{n} = \frac{\mu^{ex}}{n}$$

no local information



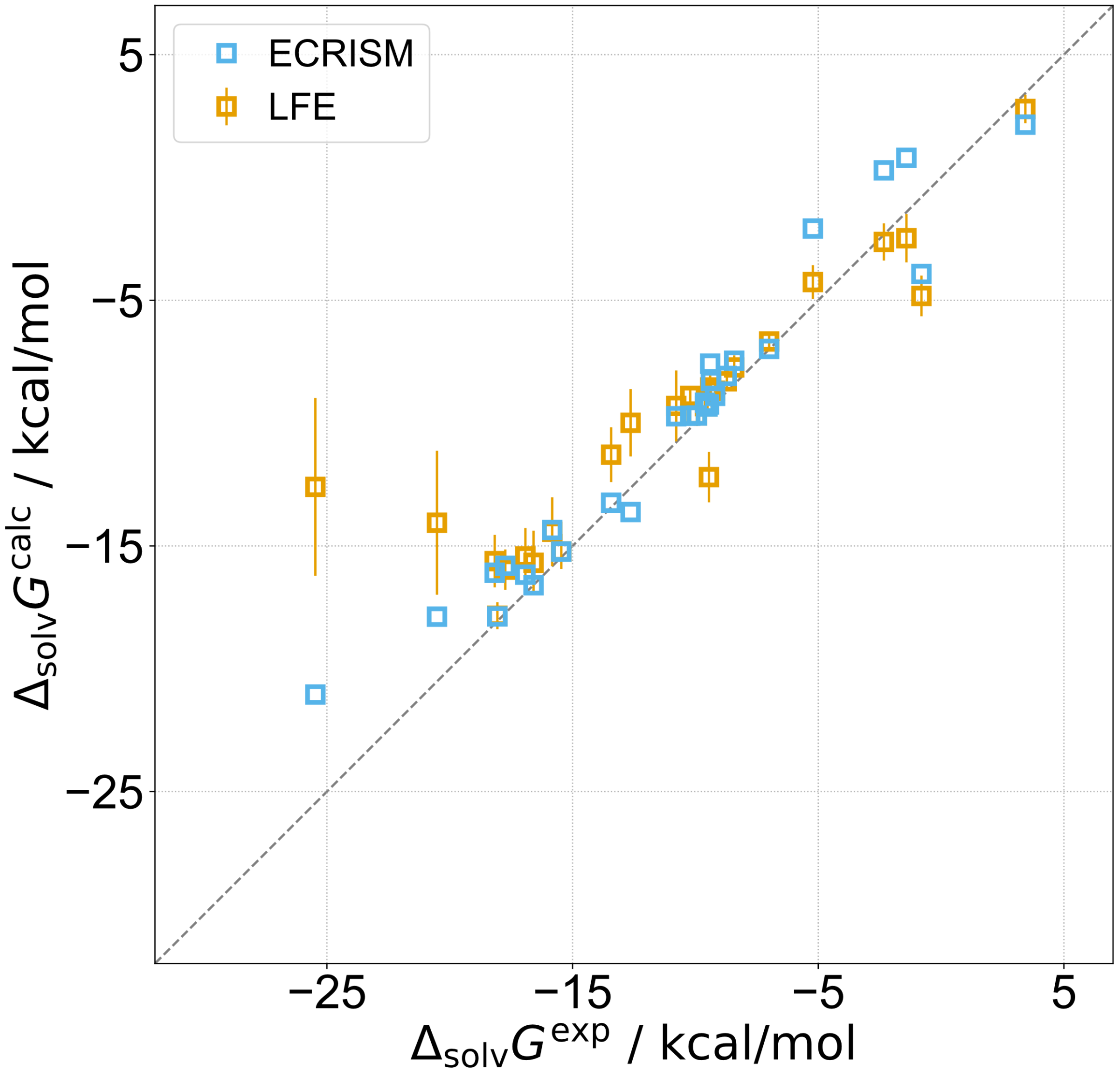
$$\mu_\alpha^{ex} \neq \mu_\alpha^{ex} \neq \mu_\alpha^{ex}$$

local information

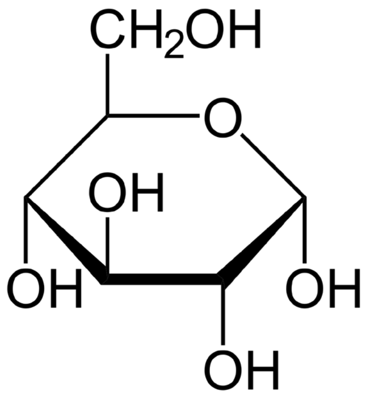
Hyperparameters for the model

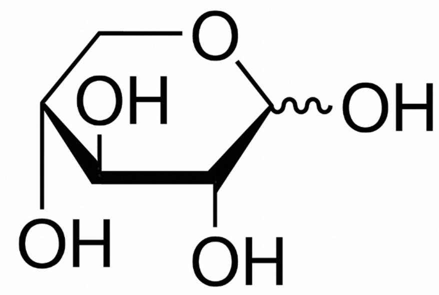
Parameter	Value
Epochs	256
Batch size	16
Learning rate	0.0005
Beta1	0.9
Beta2	0.999
Epsilon	10^-8
Weight decay	10^-8
Dimension of h	128
Number of initial passing steps	2
R hidden sizes	512,128,256,64
Message norm	Mean

SAMPL2



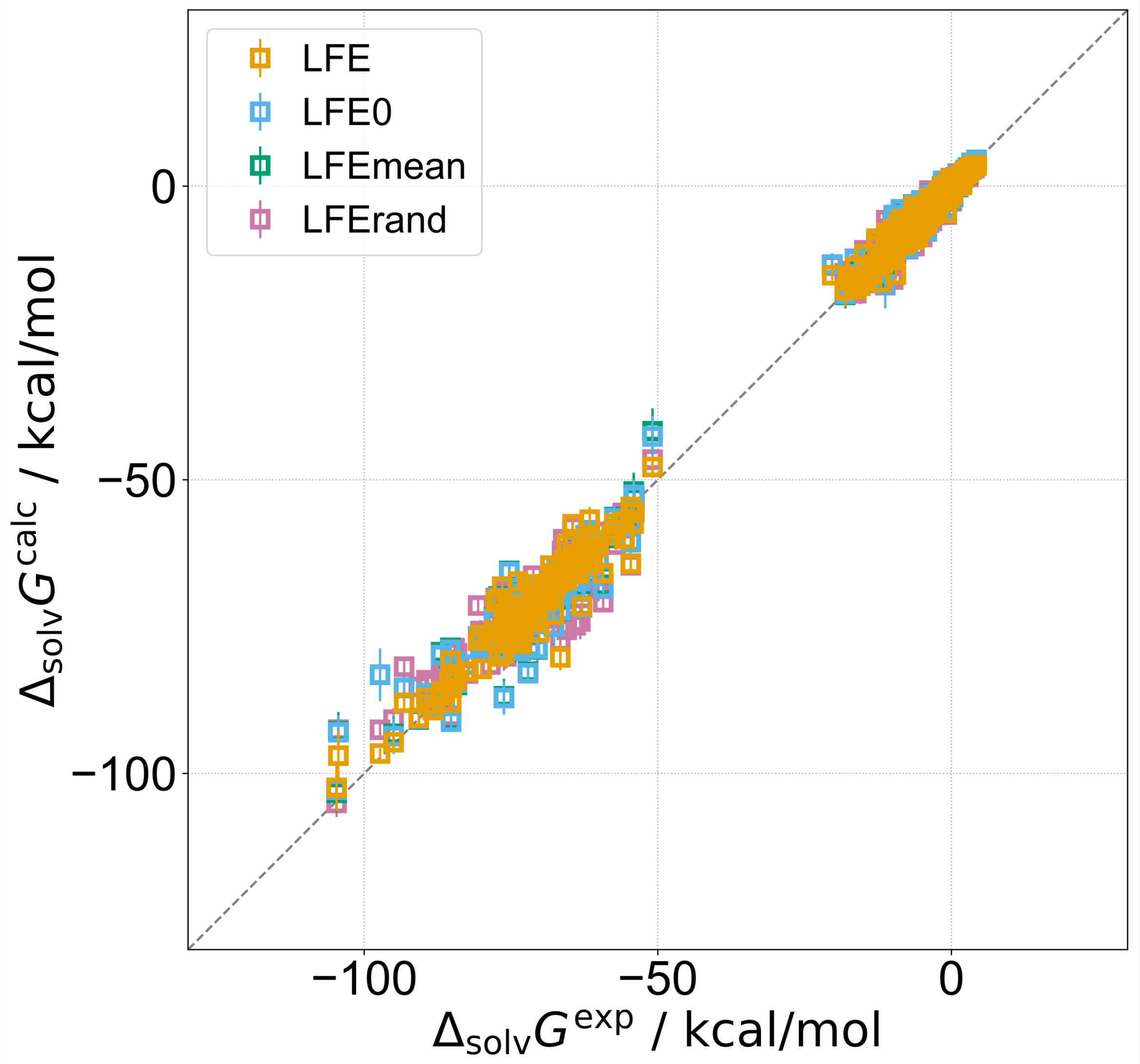
Calc	RMSE / kcal/mol	MAE / kcal/mol
ECRISM	1.68	1.24
ML	3.09	1.77
ECRISM	1.42	1.07
ML	1.51	1.16





https://static.grainger.com/rp/s/is/image/Grainger/45ZF16_AL01
https://upload.wikimedia.org/wikipedia/commons/thumb/c/c6/Alpha-D-Glucopyranose.svg/1200px-Alpha-D-Glucopyranose.svg.png

LFE info atomwise



Calc	RMSE / kcal/mol	MAE / kcal/mol
LFE	1.86	1.11
LFE0	2.27	1.36
LFEmean	2.27	1.36
LFErand	2.51	1.56