PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Check for updates

Cite this: DOI: 10.1039/d5cp01448d

The euroSAMPL1 pK_a blind prediction and reproducible research data management challenge

Nicolas Tielker, [a] Michel Lim, [b] Patrick Kibies, [a] Juliana Gretz, [c] Björn Hein-Janke, [d] Christian Chodun, [a] Ricardo A. Mata, *[d] Paul Czodrowski *[b] and Stefan M. Kast *[a]

The development and testing of methods in computational chemistry for the prediction of physicochemical properties is by now a mature form of scientific research, with a number of different methods ranging from molecular mechanics simulations, over quantum calculations, to empirical and machine learning models. Blind prediction challenges for these properties are regularly organized to allow researchers from academia and industry to test their methods in a fair and unbiased manner. At the same time, research data management (RDM) is still not utilized as extensively as it could be in the development and application of such models, especially in academia. In particular, the FAIR standards (Findable, Accessible, Interoperable, Reusable) can serve as guidelines for good RDM, but many models, the data used to train them, and the data they generate fall short of one, or multiple, of these standards. The goal of the first euroSAMPL pK_a blind prediction challenge was to promote and help develop good RDM standards for computational chemistry. To achieve this, the challenge was designed to rank not just the predictive performance of the models but also evaluate the adherence to the FAIR principles by cross-evaluation of the participants themselves. We here present the analysis of the blind prediction quality by their statistical metrics as well as of the cross-evaluation by a newly defined "FAIRscore". The results suggest that multiple methods can predict the pK_a to within chemical accuracy, but also that "consensus" predictions constructed from multiple, independent methods may outperform each individual prediction. Furthermore, the state of research data management in the field of computational chemistry is discussed, and suggestions for future improvements developed.

Received 15th April 2025,
Accepted 11th July 2025

DOI: 10.1039/d5cp01448d

rsc.li/pccp

Open Access Article. Published on 02 September 2025. Downloaded on 9/11/2025 8:59:23 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Introduction

With the rampant development of computational chemistry models and algorithms, keeping an overview of their capabilities and/or limitations can be a strenuous task. Not all methods are developed with the same chemical space in focus, nor are they tested with the same benchmarking data. Comparative performances of simulation software are rarely provided, despite the clear advantages that these community efforts bring forth. This includes verifying the reproducibility across different codes,[1] as well as asserting simulation uncertainties, a practice that is still somewhat neglected in molecular modeling.[2] Different predictors need to be compared on the same footing with a shared pool of data. On top, appropriate reference data needs to be provided, preferably with a well-defined experimental observable[3] and enough statistics for uncertainty quantification.

In this context community-organized blind-challenges play a very special role. Not only do they provide the very much needed curated pool of data, they create the conditions for an unbiased test of computational protocols. Every method has its unique switches and knobs. In the case of electronic structure theory there is the choice of chemical model (theory level for structure optimizations, solvent model, *etc.*), for machine learning approaches the training data, model layout, and hyperparameters. With the *a priori* knowledge of the target quantity any method can be adjusted, giving an unfair advantage to protocols with the most flexible parametrizations. Only by depriving the predictors of their targets can one truly assess their predictive power.

To achieve such an unbiased assessment of computational model performance various blind prediction challenges have

[a] *Department of Chemistry and Chemical Biology, TU Dortmund University, Otto-Hahn-Straße 4a, 44227 Dortmund, Germany. E-mail: stefan.kast@tu-dortmund.de*
[b] *Department of Chemistry, Johannes Gutenberg University Mainz, Duesbergweg 10-14, 55128 Mainz, Germany. E-mail: czodpaul@uni-mainz.de*
[c] *Faculty of Chemistry and Biochemistry, Ruhr University Bochum, Universitätsstraße 150, 44801 Bochum, Germany*
[d] *Institut für Physikalische Chemie, Georg-August University of Göttingen, Tammannstraße 6, 37077 Göttingen, Germany. E-mail: ricardo.mata@chemie.uni-goettingen.de*

Open Access Article. Published on 02 September 2025. Downloaded on 9/11/2025 8:59:23 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

emerged over the past decades. They share the common characteristic that the modeling community is asked to solve a certain task, the prediction of some experimentally measurable quantity, given only some molecular or system description and details about the experimental setup. Only after the previously announced challenge run time has ended experimental data are revealed to allow for statistical evaluation of model performance and ranking of different methodologies. Repeating such challenges on a single type of quantity then facilitates a historical perspective on the further development of computational models which is *per se* interesting. For instance, when the historical trend of a typical statistical metric such as the root mean squared error (RMSE) between experimental data and theoretical predictions plateaus and converges at some finite non-zero value, one can ask whether this is an indication of the limiting experimental uncertainty or of technically insurmountable difficulties to improve models further within limited resources.

The history of blind challenges for the simulation of molecular systems is fairly clear-cut. Periodic challenges include the Cambridge Structure Prediction (CSP) blind test, which goes back to 1999.[4] In this latter challenge computational predictions are assessed on a set of unpublished molecular crystals. The Critical Assessment of methods of protein Structure Prediction (CASP) provides an analogue for proteins[5] with the fourteenth edition held in 2020[6] playing a pivotal role for the recognition of deep learning methodologies (namely the Alphafold2 model[7]). Individual experimental blind challenges have also been recently pushed forward by smaller groupings,[8–12] but in such cases it is hard to keep the same continuity and level of visibility as CSP or CASP. The field of drug discovery has benefitted from – now discontinued – grand challenges (GC) organized by the D3R (drug design data resource)[13] and its predecessor, the community structure-affinity resource (CSAR).[14] A common trait across these initiatives, small or large, is the availability of experimentalists to not only conduct the necessary experiments, but also to patiently wait for the challenge to be concluded before publishing their results. This can take between months and years, depending on how fast the data analysis can be carried out and/or the participants are able to provide all the needed material for publication. Such a high effort can, however, turn out to be highly valuable, as has been demonstrated by the successful completion of the CACHE (critical assessment of computational hit-finding experiments) challenge #2.[15,16]

Another long-running challenge series is the statistical assessment of the modeling of proteins and ligands (SAMPL).[17–19] These are aimed at a critical assessment of the predictive power of computational protocols in different facets of rational drug discovery. This includes experimentally determined quantities such as binding affinities, hydration free energies, partition and distribution coefficients, with the targeted observables depending on the specific edition. In an effort to keep the community action alive and growing, new promoters for the challenge came together and created an extension, this time with the experiments being carried out in the European region, therefore coined as euroSAMPL.

The first euroSAMPL blind prediction challenge (“euroSAMPL1” in what follows) picked up an earlier target also addressed during SAMPL6–8, namely an investigation of the ability of computational methods to predict acidity constants (pK_a) for drug-like small molecules. Our primary goal was to define a set of chemically diverse, yet well-characterized and controlled compounds in the sense that only a single macroscopic transition, *i.e.* change of charge, was experimentally observed in the pH range 2–12, and for which we expected dominance of only a single tautomer in each charge state according to our own calculations. This way, we hoped to attract participants from very diverse modeling communities, ranging from atomistic, quantum-mechanical (QM) methods up to empirical rule-based and machine learning approaches, as only the macroscopic pK_a values had to be predicted without explicit reference to ensembles of coupled charge and tautomer (so-called microstate) transitions, as was required starting with the SAMPL7 challenge.[20]

This challenge design allowed for very diverse methods and, as a consequence, very different formats of primary raw data from which a single macroscopic quantity is derived. Therefore, blind prediction challenges also represent an ideal environment to test and foster adherence to modern standards of research data management (RDM). Making research data FAIR[21] (Findable, Accessible, Interoperable, Reusable) is an increasingly important requirement for research groups, scientific journals, and funding organizations, and significant progress has been made by taking advantage of the increasing digitalization of research data. Furthermore, good scientific practice demands that research data is published in a way that makes it reproducible. The reproducibility of computational chemistry data using only the information in a given journal article and its supporting information is vital for other researchers to easily verify and use newly developed methods. For the combination of FAIR data with data reproducibility standards to make RDM even “fairer”, we choose the acronym FAIR+R. The relevance of adding “reproducibility” to the FAIR principles has also independently been recognized by others.[22] This includes methods such as the automated or manual annotation of generated research data with relevant author- and domain-specific metadata, persistent storage in suitable repositories accessible to other researchers, and the transparent and – as the ultimate goal – fully automated[23] analysis of raw data to generate the chemically relevant information.

In Germany, the NFDI4Chem[24] is a consortium of the “Nationale Forschungsdateninfrastruktur” (NFDI, National Research Data Infrastructure) responsible for developing sustainable RDM standards and infrastructure for both experimental and theoretical fields in chemistry. One of the specific goals of NFDI4Chem is the design of use cases that allow for testing the usage of RDM tools, and acceptance and adherence to RDM standards in the community.[25] The euroSAMPL challenge was designed as such a use case by requiring participants to not only submit the target predictions but to also describe the methodology, including provision of underlying raw data, in a maximally transparent and reproducible format. In order to evaluate correspondence to FAIR+R principles participants

Open Access Article. Published on 02 September 2025. Downloaded on 9/11/2025 8:59:23 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

were asked after the challenge had finished to anonymously peer-evaluate the submissions of all other participants using a standardized questionnaire. The availability of prediction metrics comparing theory and experiment as well as the resulting "FAIR-scores" allowed for ranking and discussing submissions according to model and RDM quality, and their combination. This way, and with an outlook to future challenges, we intend to continually raise the bar simultaneously for both, model development and RDM standards in the computational chemistry field, expecting this challenge design to also stimulate progress toward generally accepted community-specific metadata formats.

The present paper describes the setup and timeline of the challenge including the rationale behind the choice of the systems, covering experimental details and results of our own preliminary calculations. Submissions and their evaluations are discussed, followed by analysis and interpretation of resulting model prediction metrics and FAIRscores. Insights and perspectives for future challenges conclude our report.

Challenge design

Compound preparation and measurements

The initial collection of 229 compounds was obtained from the research group of Ruth Brenk at the University of Bergen. These compounds were purchased from Otava Chemicals as part of a fragment screening library. All molecules contained at least one aromatic ring and exhibited limited flexibility, with fewer than four non-terminal rotatable bonds.

The measurements were conducted on a Sirius T3 instrument from Pion.[26] The pK_a determination was carried out in UV-metric mode, utilizing a 10 mM DMSO solution, with 5 μL of the sample prepared in 25 μL of phosphate buffer, to maintain accurate pH control throughout the titration. The analyses were conducted under argon flow at a temperature of 25.0 °C, with an ionic strength adjusted to 0.15 M using KCl in water or water/cosolvent mixtures, respectively. Each measurement was performed in triplicate within the same vial. Multiple sets of triplicate measurements for 2–4 times for the same compound were arithmetically averaged to determine the final pK_a values.

For poorly soluble compounds (*i.e.* all except **euroSAMPL-2**, **-5**, **-11**, **-13**, **-14**, **-15**, **-17**, **-19**, **-20**, **-22**, and **-27**), methanol–water mixtures adjusted to 30, 40, and 50 vv-% were used to increase solubility for the measurements over the entire experimental pH-range. From Yasuda–Shedlovsky extrapolation performed by the analytics software the pK_a values obtained in these methanol–water mixtures were then extrapolated to determine the aqueous pK_a. Based on previous measurements in various buffer/cosolvent mixtures we expect that the residual DMSO amount does not affect the aqueous equilibrium.[27] The experimental pK_a values ranged from 2.9 to 9.5.

The experimental data, presented as macroscopic pK_a values, did neither reveal which group was predominantly titrated, nor the identities of the associated macrostates (total charge), nor contributing microstates (tautomers). Additionally, the data provided no information about the charge states of the protonated and deprotonated species corresponding to each macroscopic pK_a.

The initial set of molecules were screened based on the quality of the pK_a measurements conducted on the Sirius T3. Measurements with missing datapoints, solvation issues or poor fits were excluded from the dataset. The remaining molecule representations were standardized using RDKit (version 2021_09_2) by removing salts, neutralizing charges and generating canonical SMILES representation.[28] Possible tautomers were enumerated using OpenEye's QUACPAC software (version 2.1.3).[29] Subsequently, these tautomers were cross-referenced with a comprehensive literature dataset compilation to ensure that these specific molecules had not been previously measured. The literature database comprised of different datasets, integrating data from multiple sources, including public databases such as ChEMBL,[30] DataWarrior,[31] datasets from the Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL) challenges,[20,32] experimental measurements from our laboratory, and data extracted from various publications.[33–40] The filtered data set was further processed by removing molecules with more than one stereocenter and those exhibiting more than one measured pK_a value. Stereocenters were identified using RDKit. After all filtering steps the final set of 35 euroSAMPL1 challenge molecules is depicted in Fig. 1 along with their experimental pK_a value; Fig. 2 shows molecular specification distributions.

The dominant tautomers and protonation states for our own reference calculations were generated based on pK_a values and underlying microstates predicted by ChemAxon Marvin (version 21.20).[41]

Computational details of reference calculations

By the term "reference" we mean our own set of pK_a calculations conducted before the challenge started. These were performed to ascertain by an orthogonal method that, within the experimental pH range 2–12, all compounds selected had only one macroscopic pK_a value dominated by a single microscopic transition among the microstates provided by the empirical tools described above. We followed the methodology refined during the participation of some of the authors in the SAMPL6 and SAMPL7 pK_a prediction challenges,[42,43] where EC-RISM reached RMSEs with respect to the experimental reference of 1.13 and 0.76 pK units, respectively.

3D geometries for the reference calculations were generated as follows: the SMILES string for each microstate of each compound was used to generate initial structures with RDKit's EmbedMultipleConfs module.[44] In accordance with our usual workflow, 50 conformations were generated for all microstates, as the number of rotatable bonds was smaller than 7 for the entire set of compounds.[43] These initial structures were then preoptimized with the sander utility of AMBER20, using an ALPB solvent representation with a fixed dielectric constant of 78.5.[45,46] Following this preoptimization the structures were pruned by removing all structures at least 5 kcal mol^{-1} higher in force field energy than the minimum for that microstate, and

Open Access Article. Published on 02 September 2025. Downloaded on 9/11/2025 8:59:23 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

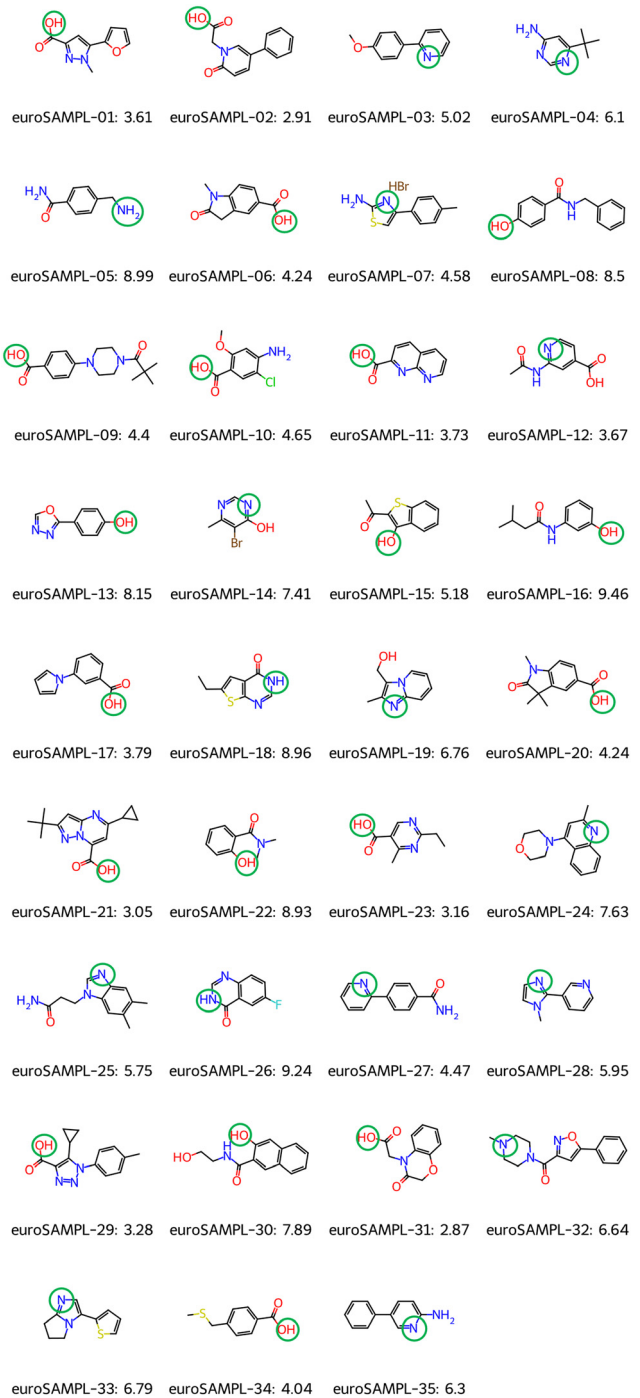

Fig. 1 euroSAMPL1 compounds and their experimentally measured pK_a values. The group at which (de-)protonation predominantly occurs, as predicted by ChemAxon Marvin, is marked with a green circle.

then clustered with a distance criterion of 0.5 Å, starting with the lowest energy conformation as the first cluster representative.

The remaining cluster representatives were further optimized at the B3LYP/6-311+G(d,p) level of theory[47,48] with the IEFPCM solvation model for water with default settings, using Gaussian 16 Rev. C.01 with tight convergence criteria, the default pruned "ultrafine" grid, and explicitly computing the force constants after converging the first SCF iteration.[49] The optimized

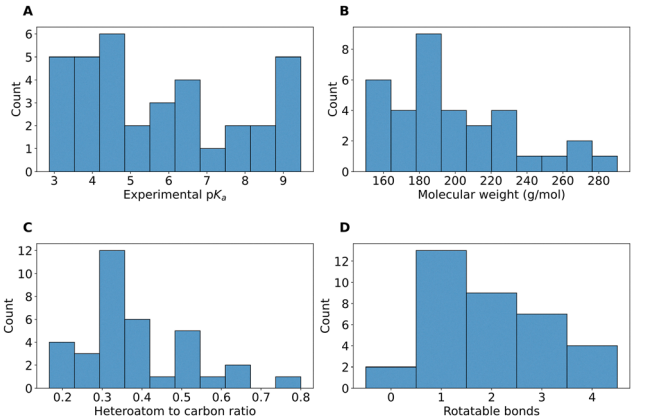

Fig. 2 Distribution of molecular specifications of the final set of euroSAMPL1 compounds.

structures were again clustered without energy cutoff and the same distance criterion, and up to five of the remaining lowest energy cluster representatives were used in subsequent calculations with the "embedded cluster reference interaction site model" (EC-RISM).[50] EC-RISM is based on a combination of the three-dimensional "reference interaction site model" (3D RISM) integral equation theory with quantum-mechanical (QM) calculations of the electronic structure.

EC-RISM calculations were conducted at the MP2/6-311+G(d,p) level of theory, using Gaussian 09 Rev. E.01 for the QM part of the calculations, as consideration of electron correlation has turned out to be essential for predicting accurate pK_a values by EC-RISM.[42,43] The radius of the electrostatic potential for bromine atoms was set to 1.3 Å. During the 3D RISM calculations, a cubic grid of 128^3 points with a spacing of 0.3 Å was employed and the solvent represented by our modified SPC/E model[51,52] with the PSE2 closure,[53] while the solute was represented by the GAFF 1.7 force field's Lennard-Jones parameters for the non-electrostatic solute–solvent interactions.[54] Electrostatic contributions were computed directly from the wave function, as outlined in ref. 43.

The molecular and tautomer energies were calculated by Boltzmann-weighting all Gibbs energies of the same ionization state, or all Gibbs energies of the same tautomer, respectively, from which the macroscopic pK_a values were derived.[43]

Challenge setup and timeline

Development of the technical infrastructure for running the first euroSAMPL blind prediction challenge began with the setup of the official challenge GitLab repository[55] that served as the central hub for the participants. Here, the initial challenge information and the compounds to be predicted were published on 2024-01-30, and any updates were published there afterwards. After the challenge, the submitted data and a preliminary analysis of the results were made available to the public. As GitLab does not provide persistent identifiers, we publish the data material also in the repositories TUDOdata and RADAR4Chem (see Data Availability statement below) as of 2025-03-14. Additionally, a qmbench.net[56] instance was created

Open Access Article. Published on 02 September 2025. Downloaded on 9/11/2025 8:59:23 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

on 2024-02-19, to serve as a web interface that allowed challenge participants to upload their results, the metadata, and optionally the raw data from their calculations. At the same time this was used to verify the completeness of the submission, including mandatory metadata fields, and the correctness of the formatting. The qmbench.net submission platform also gave each participant the option to upload one zip-archive of the raw data used to calculate the pK_a values with their method. The submission portal was closed on 2024-05-10, and after a brief, manual review to check for errors during the process, the prediction challenge results were published on GitLab, see Fig. 3.

At the same time a questionnaire about the metadata and raw data was sent to the participants, to allow them to evaluate every submission except for their own. Most participants used this opportunity, leading to each submission being evaluated by 6 or 7 peers. The results of this metadata questionnaire were combined with the prediction results and published to GitLab for the three best-performing submissions on 2024-06-11. The cross-evaluation of the metadata fields and the submitted raw data were part of the FAIR+R strategy underlying the euroSAMPL1 blind prediction challenge. Some metadata had already been collected in earlier SAMPL challenges, but here the goal was to formalize the process of collecting author-specific metadata and extend it to gathering community suggestions for domain-specific metadata that are necessary to describe their calculations and make them reproducible for other researchers. For this reason the author-specific metadata were mandatory and identical for all participants, and were selected in analogy to commonly used metadata standards, such as the Dublin Core Metadata Set and the DataCite Metadata Schema.[57,58] The domain-specific metadata on the other hand depend strongly on the specific method used, and often even differ depending on the software used to implement it. In the absence of a ready-made solution, the participants were tasked with describing, *e.g.*, the software packages and settings used for their calculations in as deep detail as necessary for other researchers to reproduce the calculations.

There were no other constraints on the type or size of the computational raw data submitted by the participants, though in practice very large submissions would have required special considerations due to limitations of the transfer protocol. The raw data could span from unstructured collections of input and log files to structured and annotated tables of energies used for the pK_a calculations, including the scripts used for this. And while none of the participants chose to do so, it would have been possible to submit one "ranked" submission (meaning the only submission entering the final score) together with additional "unranked" submissions (for which no FAIRscore would have been awarded, yet the submission's metrics would have been provided) utilizing different methods, or variations of the same method, for the pK_a prediction. A detailed graphic depicting the challenge infrastructure design is shown in Fig. 3.

The optional metadata fields and raw data formed the basis for determining the submissions' FAIRscores. It was decided early on to let the challenge participants evaluate each other's submissions with a questionnaire using Google Forms. In short, the participants were given four statements on the "findability", "interoperability", "reusability", and "reproducibility" of the meta- and research data. The research data's and metadata's relative "accessibility" was not explicitly evaluated, as the access criteria were identical for all submissions stored in the GitLab euroSAMPL challenge repository.[55] These questions, which had to be evaluated on a scale from 1 (fully agree) to 6 (fully disagree), were:

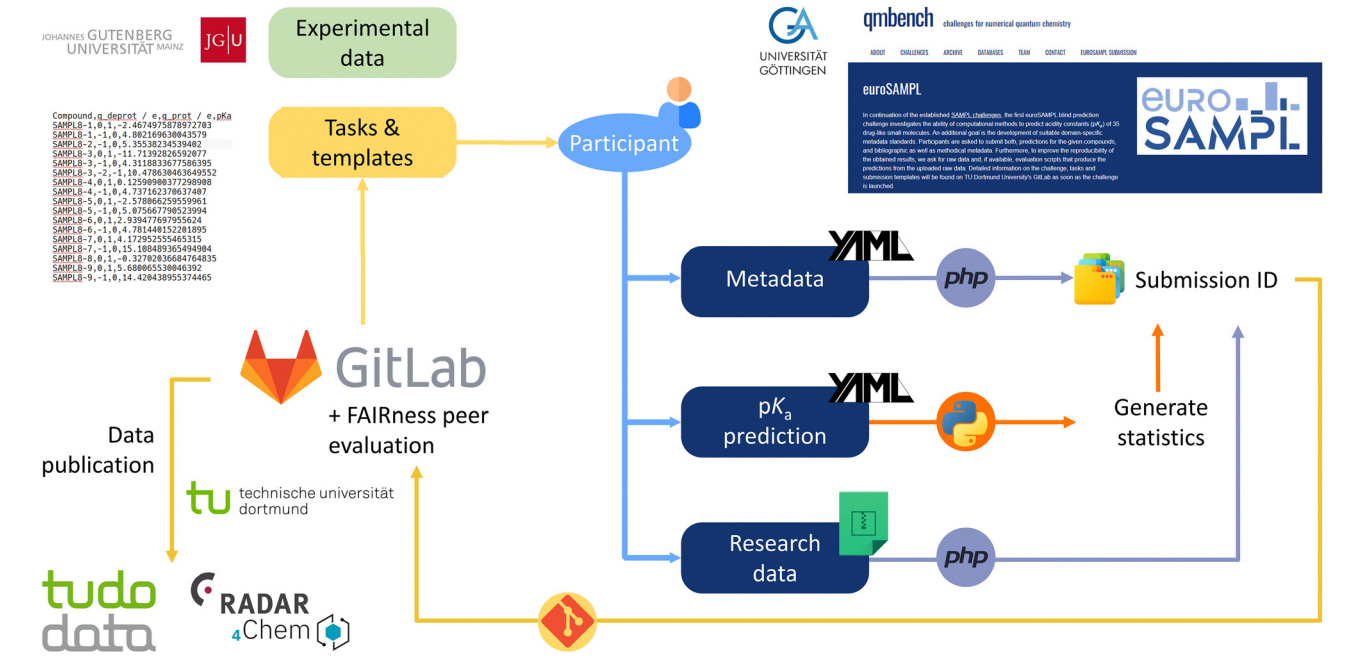

Fig. 3 Challenge infrastructure implemented for the first euroSAMPL challenge.

Open Access Article. Published on 02 September 2025. Downloaded on 9/11/2025 8:59:23 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

• The metadata field names are understandable and informative for the general audience (interoperability).

• The supplied metadata is sufficient to set up comparable calculations as were used to generate the predictions (reusability, reproducibility).

• You would use the metadata field names of the submission to search for the data in repositories that offer free-text search (findability).

• The submitted raw data and documentation are sufficient to comprehend and to enable reproduction of the predictions (reusability, reproducibility).

While these four questions alone certainly do not cover the full breadth of the FAIR principles, they were intended to let the participants evaluate the submissions' FAIRness from the point of view of experts in theoretical chemistry, not in research data management. This would allow them to take a broader perspective when considering the questions, without requiring in-depth knowledge of the FAIR criteria and their specific definitions. Because all participants had to evaluate all other participants, the average, relative evaluation of the submissions was assumed to be consistent. The FAIRscore itself was then defined as a normalized value between 0, corresponding to an average evaluation of 1.00 (fully agree), and 1, corresponding to an average evaluation of 6.00 (fully disagree).

To facilitate a combined ranking of FAIRness and prediction quality, the prediction RMSEs were also normalized to a value between 0 and 1, but due to the theoretically unbounded nature of the RMSE these values were instead mapped to the lowest RMSE of all ranked submissions, defined as an RMSEscore of 0, and the highest RMSE, defined as an RMSEscore of 1. The average of the two individual scores was then used for the final, combined ranking metric.

Results and discussion

EC-RISM reference calculations

As outlined before, our reference was designed to test for any major discrepancies between a well-established computational method of pK_a prediction and the experimentally detected values, as well as to identify compounds with more than one populated microstate for any of the relevant protonation states. The calculations were conducted prior to the challenge and yielded acidity constants in good agreement with the experimental values for the final set of 35 compounds, which exhibited an RMSE of 1.107, in line with results of earlier SAMPL challenges. This way, the experimental data can also be viewed as validated because outliers between EC-RISM and experiment would have hinted at an experimental issue due to the consistently found agreement between EC-RISM and experiment in the past SAMPL6 and SAMPL7 pK_a challenges.

As the goal of the euroSAMPL challenge was to compare the performance of different pK_a prediction methods without the additional complications arising from having to consider multiple microstates in the same protonation state, the detection of additional tautomers would have led to the exclusion of such compounds. The calculated populations and relative free energies for the originally generated microstates are shown in Table 1.

Some generated microstates of the challenge compounds systematically interconverted during the QM optimization, *i.e.*, every conformation of that microstate was optimized into a different microstate by transferring a proton. Because these tautomerizations during QM optimization always convert higher energy microstates into lower energy microstates, this systematic behavior implies that the initial microstate is not significantly populated in solution and will have no effect on the pK_a prediction.

For one of the compounds, **euroSAMPL-14**, EC-RISM suggested a microstate only 1.42 kcal mol^{-1} less favorable than the main, neutral microstate T0. However, because deeper investigation using the ChemAxon Chemicalize application did not confirm any presence of this additional neutral microstate, and ignoring the microstate would only shift the calculated pK_a by approximately 0.04 for EC-RISM, we decided to retain the compound as part of the dataset. For the remaining microstates, the energies calculated with EC-RISM yielded populations of less than 0.5%, with an energy difference to the next stable tautomer of at least 3 kcal mol^{-1} so that the energetic contribution to an acidity constant calculated with or without it would be completely negligible, given the experimental and methodical uncertainties known from previous pK_a prediction challenges.[20,32]

Challenge results: prediction quality

After closing of the submission portal, the submissions were analyzed and published on the challenge GitLab repository,[55] using the unique identifiers listed in Table 2 to identify them. In this work, we will refer to the submissions by numbers from (1) to (11) or the method name for clarity.

Automated analysis of the results (see Table 3) already showed that despite the experimental measurements revealing only a single protonation state transition for each compound, multiple participants had found and submitted additional pK_a values within or near the experimental range. This was an intended feature of the challenge design to avoid situations in which participants would have been forced to choose between two different protonation state transitions, *e.g.* from charge −1 to 0 and charge 0 to 1. For these submissions this occasionally led to a difference between our analysis schemes, either using only the "first" submitted pK_a value for each compound, or using the "best", *i.e.* the one closest to the experimental value, to generate the statistics. For the euroSAMPL1 challenge the instructions had specified that the "first" submitted value should be the one that the participants considered most likely to be the one measured experimentally.

It has been argued that the "best" matching approach is always the appropriate one,[59] because if the computational method predicts multiple protonation state transitions within the experimental range, the researcher's choice of one over the other is, to a degree, arbitrary. A rational decision can only be made if one prediction is significantly further away from the limit of the experimental range than the other one, even when

 This journal is © the Owner Societies 2025

Open Access Article. Published on 02 September 2025. Downloaded on 9/11/2025 8:59:23 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Table 1 Populations (in %) and relative Gibbs free energies (in kcal mol^{-1}) calculated using EC-RISM for the individual microstates generated by ChemAxon Marvin for each of the euroSAMPL challenge compounds. Relative Gibbs free energies of tautomerization were calculated with respect to the lowest energy microstate in a given protonation state, *i.e.* 0.00 indicates the lowest energy microstate. Microstates designated in brackets converted to the microstate before the brackets during QM optimization

Compound	Microstate	Charge	Population	ΔG_{taut}	Compound	Microstate	Charge	Population	ΔG_{taut}
euroSAMPL-1	T0	0	100.00	0.00	**euroSAMPL-19**	T0	0	100.00	0.00
	T1	−1	100.00	0.00		T1	1	100.00	0.00
euroSAMPL-2	T0	0	100.00	0.00	**euroSAMPL-20**	T0	0	100.00	0.00
	T1	−1	100.00	0.00		T1	−1	100.00	0.00
euroSAMPL-3	T0	0	100.00	0.00	**euroSAMPL-21**	T0	0	100.00	0.00
	T1	1	100.00	0.00		T1	−1	100.00	0.00
euroSAMPL-4	T0	0	100.00	0.00	**euroSAMPL-22**	T0	0	100.00	0.00
	T1	1	100.00	0.00		T1	−1	100.00	0.00
euroSAMPL-5	T0	0	100.00	0.00	**euroSAMPL-23**	T0	0	100.00	0.00
	T1	1	100.00	0.00		T1	−1	100.00	0.00
euroSAMPL-6	T0	0	100.00	0.00		T2[a]	1	100.00	0.00
	T1	−1	100.00	0.00	**euroSAMPL-24**	T0	0	100.00	0.00
euroSAMPL-7	T0	0	100.00	0.00		T1	1	100.00	0.00
	T1	1	100.00	0.00	**euroSAMPL-25**	T0	0	100.00	0.00
euroSAMPL-8	T0	0	100.00	0.00		T1	1	100.00	0.00
	T1	−1	100.00	0.00	**euroSAMPL-26**	T0	0	0.18	3.74
euroSAMPL-9	T0	0	100.00	0.00		T1	0	99.82	0.00
	T1	−1	100.00	0.00		T3	0	0.00	7.64
euroSAMPL-10	T0	0	100.00	0.00		T5[a]	−1	100.00	0.00
	T1	−1	100.00	0.00		T2	1	100.00	0.00
euroSAMPL-11	T0	0	100.00	0.00		T6	1	0.00	37.17
	T1	−1	100.00	0.00	**euroSAMPL-27**	T0	0	100.00	0.00
euroSAMPL-12	T0	0	100.00	0.00		T1	1	100.00	0.00
	T1	−1	100.00	0.00	**euroSAMPL-28**	T0	0	100.00	0.00
euroSAMPL-13	T0	0	100.00	0.00		T1	1	99.66	0.00
	T1	−1	100.00	0.00		T2	1	0.34	3.36
euroSAMPL-14	T0	0	91.49	0.00		T3[a]	2	100.00	0.00
	T1	0	8.30	1.42	**euroSAMPL-29**	T0	0	100.00	0.00
	T3	0	0.21	3.60		T1	−1	100.00	0.00
	T2(T5,T6)	−1	100.00	0.00	**euroSAMPL-30**	T0	0	100.00	0.00
euroSAMPL-15	T0	0	99.98	0.00		T1	−1	100.00	0.00
	T2	0	0.02	5.09	**euroSAMPL-31**	T0	0	100.00	0.00
	T3	0	0.00	17.68		T1	−1	100.00	0.00
	T6	0	0.00	15.51	**euroSAMPL-32**	T0	0	100.00	0.00
	T1	−1	100.00	0.00		T1	1	100.00	0.00
	T4	−1	0.00	14.61	**euroSAMPL-33**	T0	0	100.00	0.00
	T5	−1	0.00	15.64		T1	1	100.00	0.00
euroSAMPL-16	T0	0	100.00	0.00	**euroSAMPL-34**	T0	0	100.00	0.00
	T1	−1	100.00	0.00		T1	1	100.00	0.00
euroSAMPL-17	T0	0	100.00	0.00	**euroSAMPL-35**	T0	0	100.00	0.00
	T1	−1	100.00	0.00		T1	1	100.00	0.00
euroSAMPL-18	T0	0	99.99	0.00					
	T1	0	0.00	5.93					
	T2	0	0.00	6.72					
	T4	−1	100.00	0.00					

[a] States that were not included because the resulting macroscopic EC-RISM-predicted pK_a values were outside of the experimental range.

Table 2 ID, submission unique identifier on GitLab, method name given in the submitted metadata file, and method class inferred from the submitted metadata file for the nine submissions (1–9), the unranked EC-RISM calculations conducted before the challenge (10), and the Null hypothesis, assigning each compound the same pK_a value of 7.00 (11). ML refers to "pure" machine learning methods, QM means quantum mechanics-based calculations, possibly augmented by a linear correction, while QM + ML refers to QM enhanced by ML models

ID	Submission	Method name	Method class
(1)	0x4cb7101f	SP1	ML
(2)	0x4a6c0760	r2SCAN-3c/DRACO+ML	QM + ML
(3)	0xc7960c21	CBio3Lab_pK_a	ML
(4)	0x4b7b06e5	BIOVIA COSMO-RS	QM
(5)	0x216604d8	QupKake	QM + ML
(6)	0x421c06f1	H_2O_DFT	QM
(7)	0x4cb00786	RIJCOSX-B3LYP-D3BJ(SMD)/cc-pV(T+d)Z	QM
(8)	0x3f2606c6	IEFPCM_MST	QM
(9)	0x541007e2	uESE	QM
(10)	reference_EC-RISM	precalc	QM
(11)	0xb8320bc2_seven	seven	Null

Open Access Article. Published on 02 September 2025. Downloaded on 9/11/2025 8:59:23 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Table 3 Experimental values, Marvin predictions used to generate initial microstates, and submitted predictions for the pK_a values of the euroSAMPL1 challenge compounds, using both "first" and "best matching for the different submissions, designated by their ID assigned in Table 2. In cases where "first" and "best" matching leads to different predictions, both predicted values are given as "first"|"best", except for the Marvin predictions and the reference calculations (10), which were conducted non-blind and are simply sorted in descending order. The full data, including additional predictions that do not lead to different predictions depending on the matching as well as experimental and method-inherent prediction errors (to be distinguished from prediction performance measured by statistical metrics summarized in Table 5) can be found in the challenge GitLab repository and the TUDOdata and RADAR4Chem repositories (see Data Availability statement)

Compound	exp.	Marvin	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
euroSAMPL-1	3.61	3.16	3.50	3.15	4.17	3.66	3.58	4.31	4.63	3.29	2.37	3.17
euroSAMPL-2	2.91	3.90	3.49	3.63	3.73	2.74	4.23	2.06	3.04	2.01	2.98	3.66
euroSAMPL-3	5.02	4.54	5.13	4.84	4.72	6.57	4.63	4.48	8.18	0.98	5.14	4.69
euroSAMPL-4	6.10	5.69	6.54	6.23	5.75	7.17	5.83	6.92	5.45	1.64	7.61	4.69
euroSAMPL-5	8.99	9.17	8.97	8.56	7.81	9.62	8.40	7.41	11.29	11.16	−5.58|8.85	8.21
euroSAMPL-6	4.24	11.64|4.12	3.91	4.40	3.98	4.61	3.98	5.82	4.15	4.24	3.58	3.19
euroSAMPL-7	4.58	4.08	4.68	2.37	5.52	6.19	11.75|3.98	7.12	1.75	−0.27	1.22	3.95
euroSAMPL-8	8.50	8.47	8.91	7.41	8.30	8.13	8.15	7.35	11.26	11.02	11.85	9.89
euroSAMPL-9	4.40	4.71	4.41	5.36	4.24	4.94	4.56	6.77	5.58	4.23	7.86	4.70
euroSAMPL-10	4.65	4.23	4.84	4.20	4.21	5.49	4.46	2.71	4.57	5.02	4.26	5.18
euroSAMPL-11	3.73	3.91	3.45	5.33	4.20	2.88	3.88	3.40	4.73	2.68	2.51	3.52|−0.33
euroSAMPL-12	3.67	3.58	3.59	3.39	4.15	3.29	4.16	3.24	1.84	2.28	−0.53	5.14
euroSAMPL-13	8.15	8.87	8.25	8.93	7.88	8.04	8.46	6.46	10.15	10.65	5.91	9.53
euroSAMPL-14	7.41	10.69	7.96	6.91	7.53	7.08	7.78	2.90	5.15	7.28	−1.23|0.75	6.84
euroSAMPL-15	5.18	5.57	6.20	6.74	6.88	7.53	4.02	2.26	8.13	9.72	2.08	8.43
euroSAMPL-16	9.46	9.24	9.40	8.80	8.90	9.21	8.98	7.87	13.34	12.11	7.13	10.25
euroSAMPL-17	3.79	3.97	3.83	3.14	3.98	4.10	4.55	3.92	3.00	3.08	2.78	4.78
euroSAMPL-18	8.96	9.88	9.46	8.76	5.47	8.67	9.31	9.21	11.15	9.61	10.73	9.10
euroSAMPL-19	6.76	5.75	7.57	6.25	5.42	6.84	5.87	4.98	5.53	4.16	19.65|6.49	6.28
euroSAMPL-20	4.24	4.11	4.10	4.24	4.05	4.33	4.06	3.10	4.18	4.26	2.75	3.18
euroSAMPL-21	3.05	3.39	3.02	3.23	4.05	3.68	3.86|2.85	2.42	4.43	0.84	0.39	4.28
euroSAMPL-22	8.93	8.18	9.37	7.94	8.86	8.82	8.26	6.46	11.52	12.06	7.97	9.67
euroSAMPL-23	3.16	3.75|2.73	3.49	3.32	4.37	3.35	3.44|3.39	2.81	2.46	2.55	0.87	3.36|−0.04
euroSAMPL-24	7.63	8.67	9.31	8.37	6.24	8.84	7.52	9.73	13.07	3.10	10.92	6.12
euroSAMPL-25	5.75	6.25	4.90	6.32	5.93	6.28	5.54	7.54	3.12	4.15	7.62	4.81
euroSAMPL-26	9.24	10.14|5.00	9.51	8.88	4.81	2.22|8.71	3.16|9.64	11.42	11.21	3.52	11.45	9.20|0.25
euroSAMPL-27	4.47	4.29	3.91	3.82	4.43	5.12	4.26	7.35	6.74	0.39	3.19	3.79
euroSAMPL-28	5.95	5.73|3.74	6.10	5.53	4.78	5.93	5.82	6.24	4.38	3.51	5.31	4.77|0.42
euroSAMPL-29	3.28	3.03	3.63	4.68	3.89	3.26	4.30|4.21	1.42	4.34	2.87	3.03	2.20
euroSAMPL-30	7.89	7.97	9.06	6.41	6.94	8.74	7.38	5.85	8.61	7.13	20.54|6.24	8.91
euroSAMPL-31	2.87	3.52	3.35	3.07	3.71	2.81	3.56	2.27	2.10	1.50	−0.69	3.65
euroSAMPL-32	6.64	6.54	7.27	7.03	7.39	6.58	6.64	8.70	9.80	4.15	−8.38	8.00
euroSAMPL-33	6.79	6.09	6.39	6.17	5.78	6.84	6.64	7.91	4.34	4.74	7.62	5.53
euroSAMPL-34	4.04	4.07	4.02	3.73	3.91	4.32	4.08	3.95	3.75	4.11	2.26	4.90
euroSAMPL-35	6.30	6.33	6.31	6.19	5.68	7.04	5.81	6.36	8.38	3.48	5.10	4.42

accounting for expected experimental and prediction errors. If this is not the case, choosing the transition detected in the experiment is based on luck. We will get back to this point below for the discussion of individual results.

On the other hand, allowing the submission of multiple pK_a values and picking the one closest to the experimental value for analysis and comparison is only a fair method if the quality of the prediction is already known to be reasonably high. This is particularly true in the case of only a single measured pK_a value, where the order of protonation states cannot be used as an additional constraint. If one considers methods that potentially exhibit high errors in their predictions, which are explicitly encouraged to participate in this kind of blind prediction challenge to identify why or for which molecules such errors occur, it is possible that the "best" predicted value stems from a different protonation state transition than the one observed experimentally. While the argument that forcing the researcher to choose one of their predictions to be the "correct" one makes the comparison based on arbitrary factors has merit, we believe that the comparison of both matching methods has advantages for the purpose of blind prediction challenges. A method that coincidentally predicts the "correct" pK_a while predicting the wrong charge states around this transition is unusable for many practical applications.

Unlike during, *e.g.*, the SAMPL6 challenge[32] there is no ambiguity that would necessitate deciding on a specific matching algorithm, as each molecule only has a single experimental pK_a value to which a single prediction must be matched.

As shown in Table 5 and Fig. 4(A), using the "first" type of matching, the three best-performing submissions are described by the authors as "SP1" (1), "r2SCAN-3c/DRACO + ML" (2), and "CBio3Lap_pK_a" (3), with RMSEs of 0.53, 0.81, and 1.21 pK units, respectively. These performances are generally in line with or even slightly better than the results of earlier SAMPL pK_a prediction challenges, where for instance the best-performing submissions achieved RMSEs of 0.68 and 0.72 during the SAMPL6 and SAMPL7 challenges, respectively.[20] The submissions with these results consist of two ML and one QM + ML model.

Utilizing the "best" predicted pK_a value instead reveals a slightly different picture. In particular, the RMSEs of submissions

This journal is © the Owner Societies 2025

Open Access Article. Published on 02 September 2025. Downloaded on 9/11/2025 8:59:23 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

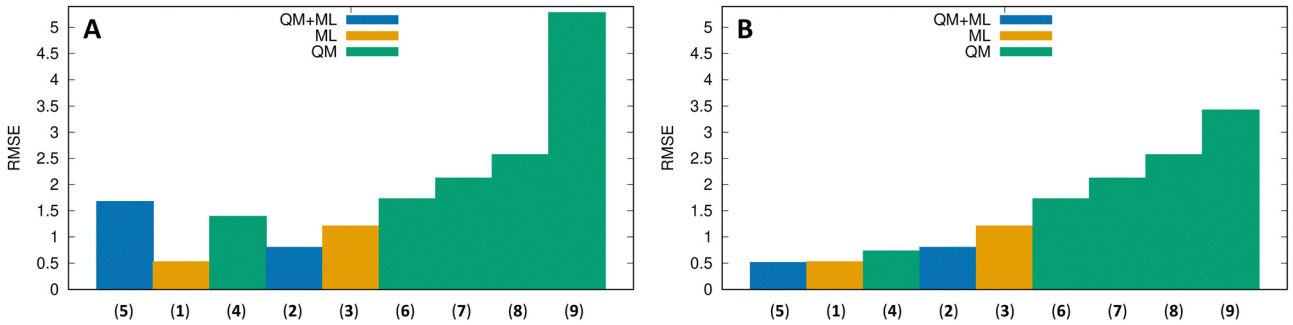

Fig. 4 Root mean square error (RMSE) of each ranked method's predictions, *i.e.* excluding reference results (10, RMSE 1.107) and Null hypothesis (11, RMSE 2.444), over the entire euroSAMPL dataset, with the method IDs and colorations set according to Table 2. Results utilizing the "first" matching approach are depicted in panel A and results utilizing the "best" matching approach in panel B, both sorted in "best" matching rank order.

"QupKake" (5) and "BIOVIA COSMO-RS" (4), a QM + ML and a QM model with linear correction, change their ranking from fourth and fifth place to third and overall best prediction, respectively. This indicates that, while the correct transitions were calculated, they were not identified as the most likely to occur within the experimental range, leading to significant deviations for a few individual challenge compounds when using the "first" matching approach. In the case of "QupKake" (5) the change in the RMSE from 1.67 to 0.51 is caused by just two compounds, **euroSAMPL-07** and **euroSAMPL-26**. In both cases the submission includes two pK_a values, 11.75 and 3.98 for the former, and 3.16 and 9.64 for the latter, while the experimentally determined values are 9.24 and 4.58, respectively. In these cases, a different heuristic for deciding on the "first" pK_a value, such as choosing the value that is farthest away from the limits of the experimental range, would have yielded the same results as the "best" matching for these two compounds. As for most other compounds this was in fact the case for this submission it would have been valuable to add the method of deciding on the "first" pK_a value to the method description in the metadata file to enable deeper investigation.

For the submission "BIOVIA COSMO-RS" (4) the difference in the RMSEs is slightly smaller, with 1.39 for "first" and 0.73 for "best" matching, and it is caused by only a single compound, again **euroSAMPL-26**. In this case the first submitted pK_a value is 2.22, and the second 8.71, and it is the only compound for which two pK_a values were submitted. Further investigation of the compound **euroSAMPL-26**, which was identified by both submissions employing "BIOVIA COSMO-RS" (4) and "QupKake" (5) as having an additional protonation state within the experimental range, revealed that our reference calculations assigned this transition a pK_a value of 0.25 (assuming an additional $-1 \rightarrow 0$ transition), well outside the experimental range.

During the 18th German Conference of Cheminformatics (**https://www.gdch.de/gcc2024**), the author of the submission "SP1" (1) presented additional compounds for which the method predicted multiple protonation state transitions within the experimental range. In light of this post-challenge analysis (to trigger this is actually one goal of organizing blind prediction challenges) we reviewed our initial micro- and macrostate set also for compounds **euroSAMPL-09**, **euroSAMPL-12**, and **euroSAMPL-28**. This led to the inclusion of a +1 state for **euroSAMPL-09**, a −2 and +1 state for **euroSAMPL-12**, and a +2 state for **euroSAMPL-28**. Corresponding EC-RISM-predicted macroscopic pK_a values are 4.19 ($0 \rightarrow +1$) for **euroSAMPL-09** (orig. 4.70 for $-1 \rightarrow 0$, exp.: 4.40), 1.83 ($0 \rightarrow +1$) and 12.08 ($-2 \rightarrow -1$) for **euroSAMPL-12** (orig. 5.14 for $-1 \rightarrow 0$, exp.: 3.67), and 0.42 ($+1 \rightarrow +2$) for **euroSAMPL-28** (orig. 4.77 for $0 \rightarrow +1$, exp.: 5.95). This means that only **euroSAMPL-09** very likely exhibits one additional transition that is fully within the experimental range and was missed in our early assessment. "SP1" (1) indeed submitted two values for **euroSAMPL-09**, for which the "first" submission correctly matched the experimental reference better. The authors also submitted two values for **euroSAMPL-28**, again "first" matching best. For **euroSAMPL-12**, only one value was submitted, in line with the EC-RISM analysis that other values lie outside the experimental pH range.

The experimental data for some of the discussed compounds like **euroSAMPL-26** also indicates signs of another protonation state, though insufficient measurement points could be collected inside the experimental pH range from 2 to 12. This is most notably also the case for **euroSAMPL-3** and **euroSAMPL-28** in the lower pH range, but in all of these cases, the potentiometric transition is only starting within the experimental range, and there is no sign of the inflection point and leveling off after the shift in the raw data. In retrospect, it might have been more unambiguous to define the experimental range investigated in the challenge more restrictively, *e.g.* as "experimental values between 2.5 and 11.5". This would have accounted for the need to see most or all of the potentiometric transition to determine an experimental value for the pK_a, whereas at a molecule's predicted pK_a, at most 50% of the species is (de-)protonated, if no other pK_a value is in close proximity. On the other hand, the "experimental range" of the potentiometric measurements ranged from a pH of 2 to 12. This knowledge should have allowed for a rational decision about the single submitted pK_a value by discounting values closer to the edge of the experimental range.

For **euroSAMPL-09** the issue is more complex: the close proximity of the predicted pK_a values makes it impossible to clearly distinguish two different protonation state transitions. The additional pK_a value should have been detected during the

Open Access Article. Published on 02 September 2025. Downloaded on 9/11/2025 8:59:23 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

pre-challenge analysis, and in that case the compound would have been removed from the dataset, but the critical microstate where the amine nitrogen is protonated had not been automatically identified. This suggests that in future challenges multiple, orthogonal approaches for the detection of protonation states and their underlying tautomers should be used to minimize the chances of this occurring. Due to the good performance of most methods on this compound, the relative ranking of the methods would not change upon removal of **euroSAMPL-09** from the analysis, even for methods as close in RMSE as "SP1" (1) and "QupKake" (5), and even absolute changes are $<1.5\%$ of the RMSE in all cases.

Beyond the relative ranking of the participating methods, there is also the question of trends in the relative performance of different methods for different compounds. This can help identify issues with individual experimental results, prompting a reinvestigation of the experimental data, and with a given method's predictive performance on certain substance classes. The euroSAMPL1 challenge compounds can be broadly divided into containing acidic carboxyl, aromatic hydroxy, and pyrimidinone functions, and a variety of basic aliphatic and aromatic amines.

Breaking down the predictions on the individual compounds reveals that, for many molecules, there are only minor variations in the predicted pK_a values. However, outliers greater than 2.5 pK units occur for five of the nine ranked submissions, namely "CBio3Lab_pK_a" (3), "H_2O_DFT" (6), RIJCOSX-B3LYP-D3BJ(SMD)/cc-pV(T+d)Z (7), "IEFPCM_MST" (8), and "uESE" (9), even when the "best" matching approach is used. This leads to significantly increased RMSEs for these even in cases where the prediction performance is good for the majority of the challenge dataset. It is noticeable that the two best-performing submissions in the "first" matching evaluation, "SP1" (1) and "r2SCAN-3c/DRACO + ML" (2) (see Table 5) also have the lowest maximum absolute error (MaxAE) between prediction and experimental value, indicating that the absence of outliers is key to good statistical performance. In fact, looking at the "best" matching statistics, with one exception where the RMSE and MaxAE values are very close (0.806/0.734 and 2.21/2.35, respectively), the order of the MaxAEs is the same as the order of the RMSEs. There is no significant clustering of the outliers for the same molecule predicted by different submissions, indicating that they are caused by methodological errors, either in the prediction itself or in the selection of microstates, not problems with the experimental setup. Furthermore, for 33 of the 35 compounds there is at least one prediction among all submissions that predicts the experimentally measured pK_a to within 0.2 pK units. The exceptions to this are **euroSAMPL-30**, where the closest prediction made by submission (5) was off by 0.51, and **euroSAMPL-15**, where the closest prediction made by submission (1) was off by -1.02. However, even in these cases, the average deviations between predictions of the ranked submissions and experimental pK_a values are 0.38 ± 1.15 and -0.83 ± 2.48, respectively, indicating only weak agreement among the different theoretical methods.

Chemically, these two compounds have in common that they are both aromatic alcohols, but a number of other compounds such as **euroSAMPL-08**, **euroSAMPL-14**, and **euroSAMPL-16** do not systematically exhibit this large of a mismatch between the predictions and the experimentally measured pK_a values.

As another matter of interest, even though the number of submissions is rather small, a synthetic submission using the Null hypothesis of a pK_a of 7.00 for all compounds, the center of the experimental range, yields an RMSE of only 2.44, better than some of the submissions. On the other hand, using the "average" predicted pK_a of all ranked submissions with their "best" matching as a prediction, would have yielded an RMSE of only 0.56 pK units, only 0.05 worse than the best-performing method "QupKake" (5). This is despite the fact that the average standard deviation of the individual predictions is as large as 1.47 pK units. Even more noticeable, restricting the average to the top 5 submissions (1)–(5) would have led to an RMSE of only 0.39, handily winning the challenge. Here, the average disagreement between methods, as measured by their predictions' standard deviations was still 0.59 pK-units. This shows that while the individual predictions taken from different methods may in some cases miss the mark, and even disagree quite significantly with each other, their average value results in a very accurate prediction.

Challenge results: FAIRscore

In accordance with the challenge guidelines, the FAIRscore resulting from the cross-evaluation of the participants will only be discussed individually for the best-scoring submissions, and the results are summarized in Table 4. The method "RIJCOSX-B3LYP-D3BJ(SMD)/cc-pV(T+d)Z" (7) was ranked first place among all participants, with a mean FAIRscore of 1.54 (normalized: 0.228). This submission had not only an exhaustive metadata file for the method used, but also an extensive raw data folder. This folder contained the output files of each calculation, in which the content of the original input file is also included, as well as a table of the electronic energies and the entire calculations that were conducted to yield the submitted pK_a values.

This exemplarily FAIR submission was followed by three equally FAIR submissions in second place, namely "SP1" (1), "r2SCAN-3c/DRACO + ML" (2), and "uESE" (9) with an identical FAIRscore of 2.00 (normalized: 0.421), a gap of 0.46 to the FAIRest submission. While the aggregate scores for these submissions are identical, the underlying individual scores, and thus the reasons for their slightly worse FAIRscores, differ. While the scores for Q1 are reasonably close together, submission "uESE" (9) scores better in reproducibility and reusability (Q2 and Q4) while scoring lower in findability (Q3). As this is the only pure physics-based QM method among these three, one needs to think about the perception of the peers when it comes to assessing "reproducibility". One might argue that the distinction between empirical and physics-based methods is related to software availability on the one hand and to data handling on the other hand. For empirical methods, the target property can potentially be obtained from the structural input directly, hence "reproducibility" hinges upon direct access to the software. In contrast, for physics-based methods several post-processing steps connect primary physical data with the target property in question. Hence, it is important that all raw data and

Open Access Article. Published on 02 September 2025. Downloaded on 9/11/2025 8:59:23 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Table 4 Individual scores for the FAIRscore evaluation of the four best-performing submissions

ID	Q1	Q2	Q3	Q4	Ø
(7)	1.17	1.67	1.67	1.67	1.54
(1)	1.50	2.00	1.67	2.83	2.00
(2)	1.43	1.86	1.86	2.86	2.00
(9)	1.67	1.67	2.33	2.33	2.00

metadata are available, ideally in combination with programs to extract the target property from raw data. In this case, primary software need not necessarily be available and a user might be content with the published information. In conclusion, when the software for empirical models is not freely available, peers could be tempted to assign lesser "reproducibility".

Submission "RIJCOSX-B3LYP-D3BJ(SMD)/cc-pV(T+d)Z" (7) also provided the most extensive raw data. Only for one other submission the same output files were provided, but in that case no processed data such as energies, or the calculation of the acidity constants from the raw energies were part of the raw data. This is most noticeable in the evaluation of Q4, where the largest gaps, ranging from 0.66 to 1.19, between the highest-ranked submission and the three runners-up occurs.

Combining the normalized results of the pK_a prediction with the normalized FAIRscores, as shown in Table 5, also shows that even submissions which do not perform well in the pK_a prediction part of the challenge can be FAIR+R, with the best FAIRscore assigned to one of the lower-ranked submissions, "RIJCOSX-B3LYP-D3BJ(SMD)/cc-pV(T+d)Z" (7), significantly improving its combined rating from seventh to fourth rank using "best" matching and third rank using "first" matching. On the other hand, the overall combined ranking still rewards a good predictive performance, with the first and second place remaining the same due to their good FAIRscore.

Conclusions, lessons learned, and perspectives

With the first euroSAMPL challenge successfully concluded, the results show that a number of different methods, ranging from empirical to quantum-mechanics based, are able to predict acidity constants to within what is usually called "chemical accuracy", *i.e.* an error of less than 1 kcal mol^{-1} which is equivalent to a pK_a difference of approximately 0.73.[60] However, one needs to conceptually distinguish between "first" and "best" submission performances, as the former measures a truly blind quality, in the absence of any *a priori* knowledge about the outcome, while the latter can only be defined after the experimental reference data have been revealed, *i.e.* including *a posteriori* knowledge. The two best-performing methods when using "best" matching, "SP1" (1) and "QupKake" (5) are able to beat this standard by a significant margin, with the method "BIOVIA COSMO-RS" (4) hitting it exactly. In this sense, a state-of-the-art performance for pK_a predictions as a result of this challenge is characterized by an RMSE of 0.5, in line with the most optimistic

Table 5 Statistical metrics, normalized RMSE, FAIRscore, and combined score for all submissions using either "first" or "best" matching for submissions with multiple predicted acidity constants for the same compound. ID refers to the submissions as defined in Table 2, RMSE, MAE, and MSE are the root mean square error, mean absolute error, and mean signed error, respectively, MinAE and MaxAE are minimum and maximum absolute errors per submission, nRMSE and nFAIR are the normalized RMSE and FAIRscore, as defined in the manuscript, and nComb is the mean value of the two. Colored squares indicate the rank of the submission in accordance with the commonly used "gold", "silver", and "bronze" for first, second, and third place, respectively. FAIRscores are shown for the best-scoring submissions only. Reference EC-RISM results (10) and those from the Null hypothesis (11) are added for completeness

ID	RMSE	MAE	MSE	MinAE	MaxAE	nRMSE	nFAIR	nComb
First								
(1)	0.529	0.379	0.214	0.01	1.68	0.100	0.421	0.261
(2)	0.806	0.632	−0.086	0.00	2.21	0.153	0.421	0.287
(3)	1.207	0.812	−0.248	0.04	4.43	0.229		
(4)	1.392	0.705	0.131	0.02	7.02	0.264		
(5)	1.672	0.779	0.016	0.00	7.17	0.317		
(6)	1.726	1.410	−0.218	0.06	4.51	0.327		
(7)	2.123	1.757	0.715	0.06	5.44	0.402	0.228	0.315
(8)	2.569	2.009	−0.945	0.00	5.72	0.486		
(9)	5.280	3.375	−0.859	0.07	15.02	1.000	0.421	
(10)	1.107	0.935	0.047	0.04	3.25	0.100		
(11)	2.444	2.142	1.276	0.21	4.13	0.153		
Best								
(1)	0.529	0.379	0.214	0.01	1.68	0.155	0.421	0.288
(2)	0.806	0.632	−0.086	0.00	2.21	0.236	0.421	0.328
(3)	1.207	0.812	−0.248	0.04	4.43	0.353		
(4)	0.734	0.519	0.316	0.02	2.35	0.215		
(5)	0.513	0.408	−0.053	0.00	1.32	0.150		0.361
(6)	1.726	1.410	−0.218	0.06	4.51	0.504		
(7)	2.123	1.757	0.715	0.06	5.44	0.620	0.228	
(8)	2.569	2.009	−0.945	0.00	5.72	0.751		
(9)	3.422	2.231	−1.175	0.07	15.02	1.000	0.421	
(10)	1.107	0.935	0.047	0.04	3.25	0.100		
(11)	2.444	2.142	1.276	0.21	4.13	0.153		

Open Access Article. Published on 02 September 2025. Downloaded on 9/11/2025 8:59:23 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

accuracy estimate of the QupKake developers derived from the analysis of their own model.[61] Only "SP1" (1) approaches this limit also in the "first" matching setup, *i.e.* under fully blinded conditions. That such a small error has been reached is even more impressive due to the design of the challenge that assured that none of the target compounds were part of the training or benchmark sets. This also confirms earlier results, as such an error margin had already been observed in previous blind prediction challenges and other practical applications in both academia and industry.[19,32,62] This observation could therefore also be viewed as proof of principle for the experimental setup in combination with the challenge design, which provide a useful platform for defining the cutting edge in terms of accuracy limit of computational prediction models.

One interesting result of the analysis is the very good performance of a synthetic "consensus" model. Using the mean prediction of many orthogonal methods appears to systematically improve the prediction quality, even though the methods' individual predictions spread significantly around the mean. Similar results have been noted before,[62] however in that case only empirical models were used for what the authors call "data fusion". The phenomenon, especially including physics-based, mixed, and pure ML/empirical models, should be tested on a larger dataset of pK_a values, as in practice this could allow researchers to predict pK_a values more accurately by using multiple fast methods to generate a consensus prediction. If this is found to be the case, its applicability to other physicochemical properties like partition and distribution coefficients or solubilities should be investigated as well.

The novelty of this challenge was the completely redesigned approach towards improved research data management, and in this domain some key insights were obtained: as a result from analyzing the peer evaluation results, the collection of meaningful metadata and raw data appears to be easier for physics-based methods where primary physical data (such as energies per structure) are automatically generated during the calculations. Derived from post-processing of primary data by a well-defined mathematical framework, the final prediction's – *i.e.* the target observable's – provenance can be described in such a way that the calculation can be reproduced by others without access to the original software that produced the primary data. Conversely, empirical and ML tools often generate only a small amount of research data in the first place, unless models, *i.e.* software along with parameters are made available. Some might only use a molecular identifier as input to generate a prediction as output; however, the models used by these methods are more complex and would benefit from increased "FAIR+Rness" in the sense of making models freely available for maximizing reproducibility. The perception of participants appears to attribute a lesser degree of reproducibility to this class of models, just because only a small amount of research data is published.

In any case, better documentation of empirical methods is possible in order to avoid the impression of using a "black box", but this demands considerable additional effort compared to physics-based methods. In a hallmark paper, Heil *et al.* defined a set of reproducibility standards for machine learning methods,[23] classifying them as either "bronze", "silver", and "gold". While the bronze standard is not too difficult to achieve from a technical perspective, requiring only that the data, models, and source code are published and downloadable, this may not always be desired by the authors of the model. A compromise could be to aim for at least fulfilling the FAIR standards for data, models, and source code, allowing for proprietary models and data to be used while improving overall research data management standards. Another option that has been investigated by the earlier SAMPL challenge maintainer and organizer David Mobley is the containerization of challenges, which would require the participants to submit a Docker container that, upon taking a pre-disclosed input, generates the prediction results.[63] In theory, this would allow researchers to keep their model confidential, while fully disclosing it to the challenge organizers.

The silver and especially the gold standard are significantly more difficult to achieve, requiring the models to be set up in a way that enables implementation of the environment and reproduction of the results in a deterministic fashion with a single click, but it should still be a goal to strive for in the long run.

Similarly, among physics-based methods, while during this challenge the input and output files of the computations as well as the analysis of the raw data to produce the final productions have been supplied by the FAIRest submission "RIJCOSX-B3LYP-D3BJ(SMD)/cc-pV(T+d)Z" (7), this should only be considered as the first steps towards truly FAIR+R RDM. By utilizing the suggestions for metadata field names supplied for the different computational methods, it should be possible to help the NFDI in developing standardized ontologies for domain-specific metadata that can be extended in the future when new programs or methods become available. Then, not only can the research data be annotated with FAIR metadata, but they can also be made more reproducible by automating and containerizing its generation.

Future euroSAMPL challenges will focus not just on the prediction of simple properties such as macroscopic acidity constants. Instead, molecular properties such as, *e.g.*, microscopic acidity constants, which may be accessible with combined NMR and potentiometric measurements, will probably a major focus. Also, problems for which ML methods are usually not specifically trained are of interest, such as, *e.g.*, temperature-dependent pK_a values or non-aqueous systems. Increased collaborations with other experimental groups could also allow for a renewed focus on blind challenges for larger systems like modeling of protein–ligand or host–guest interactions, as in many of the previous original SAMPL challenges.

One additional task that must be addressed by us as challenge organizers as well as by the wider community, is the lack of gender diversity as far as the challenge participants are concerned. For instance, despite the growing number of female PIs in the field, in both academic and industry roles, no female PI participated in this challenge. Investigating the reasons for this discrepancy and increasing our outreach to foster a truly representative environment is an important task for the development of future challenges.

Open Access Article. Published on 02 September 2025. Downloaded on 9/11/2025 8:59:23 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Author contributions

Nicolas Tielker: data curation, investigation, methodology, validation, visualization, writing – original draft, writing – review and editing, Michel Lim: data curation, visualization, investigation, Patrick Kibies: data curation, formal analysis, software, validation, visualization, Juliana Gretz: data curation, formal analysis, investigation, validation, Björn Hein-Janke: software, validation, Christian Chodun: visualization, Ricardo A. Mata: conceptualization, funding acquisition, resources, supervision, writing – original draft, Paul Czodrowski: conceptualization, funding acquisition, resources, supervision, Stefan M. Kast: conceptualization, funding acquisition, resources, supervision, writing – original draft, writing – review and editing.

Conflicts of interest

There are no conflicts to declare.

Data availability

Data for this article, including all computational and experimental raw data of our own analysis and provided by the challenge participants are available at **https://gitlab.tu-dortmund.de/kast_ccb/eurosampl/challenge**, **https://doi.org/10.17877/TUDODATA-2025-M6KPZZGR**, **https://doi.org/10.22000/dfqzn3tat216pyzy**.

Acknowledgements

This work would not have been possible without the active contribution of our challenge participants: O. Abarbanel, R. Balabin, M. Diedenhofen, R. Fraczkiewicz, A. Hellweg, G. Hutchison, E. Lichak, A. S. Paluch, P. Patel, S. A. Rodriguez, J. Uranga, A. Viayna, W. J. Zamora. The project was funded by the Deutsche Forschungsgemeinschaft (DFG) under the Chemistry Consortium of the Nationale Forschungsdateninfrastruktur – NFDI4Chem – Projektnummer 441958208 (to S. M. K., R. A. M., P. C.), and under Germany's Excellence Strategy – EXC-2033 – Projektnummer 390677874 (to S. M. K.). We also thank Tam Dieu Cao for laboratory support and the ITMC of TU Dortmund University for providing the GitLab and Dataverse (TUDOdata) instances. Last but not least, we want to express our gratitude to David Mobley as well as all previous SAMPL challenge (co-) organizers for supporting the continuation of the SAMPL idea.

References

1 K. Lejaeghere, G. Bihlmayer, T. Björkman, P. Blaha, S. Blügel, V. Blum, D. Caliste, I. E. Castelli, S. J. Clark, A. D. Corso, S. de Gironcoli, T. Deutsch, J. K. Dewhurst, I. Di Marco, C. Draxl, M. Dułak, O. Eriksson, J. A. Flores-Livas, K. F. Garrity, L. Genovese, P. Giannozzi, M. Giantomassi, S. Goedecker, X. Gonze, O. Grånäs, E. K. U. Gross, A. Gulans, F. Gygi, D. R. Hamann, P. J. Hasnip, N. A. W. Holzwarth, D. Iuşan, D. B. Jochym, F. Jollet, D. Jones, G. Kresse, K. Koepernik, E. Küçükbenli, Y. O. Kvashnin, I. L. M. Locht, S. Lubeck, M. Marsman, N. Marzari, U. Nitzsche, L. Nordström, T. Ozaki, L. Paulatto, C. J. Pickard, W. Poelmans, M. I. J. Probert, K. Refson, M. Richter, G.-M. Rignanese, S. Saha, M. Scheffler, M. Schlipf, K. Schwarz, S. Sharma, F. Tavazza, P. Thunström, A. Tkatchenko, M. Torrent, D. Vanderbilt, M. J. van Setten, V. Van Speybroeck, J. M. Wills, J. R. Yates, G.-X. Zhang and S. Cottenier, *Science*, 2016, **351**, aad3000.

2 P. Pernot, *J. Chem. Phys.*, 2022, **156**, 114109.

3 R. A. Mata and M. A. Suhm, *Angew. Chem., Int. Ed.*, 2017, **56**, 11011–11018.

4 J. P. M. Lommerse, W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Gavezzotti, D. W. M. Hofmann, F. J. J. Leusen, W. T. M. Mooij, S. L. Price, B. Schweizer, M. U. Schmidt, B. P. van Eijck, P. Verwer and D. E. Williams, *Acta Crystallogr.*, 2000, **B56**, 697–714.

5 A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis and J. Moult, *Proteins*, 2023, **91**, 1539–1549.

6 A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis and J. Moult, *Proteins*, 2021, **89**, 1607–1617.

7 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583–589.

8 K. I. Assaf, M. Florea, J. Antony, N. M. Henriksen, J. Yin, A. Hansen, Z. Qu, R. Sure, D. Klapstein, M. K. Gilson, S. Grimme and W. M. Nau, *J. Phys. Chem. B*, 2017, **121**, 11144–11162.

9 H. C. Gottschalk, A. Poblotzki, M. A. Suhm, M. M. Al-Mogren, J. Antony, A. A. Auer, L. Baptista, D. M. Benoit, G. Bistoni, F. Bohle, R. Dahmani, D. Firaha, S. Grimme, A. Hansen, M. E. Harding, M. Hochlaf, C. Holzer, G. Jansen, W. Klopper, W. A. Kopp, L. C. Kröger, K. Leonhard, H. Mouhib, F. Neese, M. N. Pereira, I. S. Ulusoy, A. Wuttke and R. A. Mata, *J. Chem. Phys.*, 2018, **148**, 014301.

10 H. C. Gottschalk, A. Poblotzki, M. Fatima, D. A. Obenchain, C. Pérez, J. Antony, A. A. Auer, L. Baptista, D. M. Benoit, G. Bistoni, F. Bohle, R. Dahmani, D. Firaha, S. Grimme, A. Hansen, M. E. Harding, M. Hochlaf, C. Holzer, G. Jansen, W. Klopper, W. A. Kopp, M. Krasowska, L. C. Kröger, K. Leonhard, M. M. Al-Mogren, H. Mouhib, F. Neese, M. N. Pereira, M. Prakash, I. S. Ulusoy, R. A. Mata, M. A. Suhm and M. Schnell, *J. Chem. Phys.*, 2020, **152**, 164303.

11 T. L. Fischer, M. Bödecker, S. M. Schweer, J. Dupont, V. Lepère, A. Zehnacker-Rentien, M. A. Suhm, B. Schröder, T. Henkes, D. M. Andrada, R. M. Balabin, H. K. Singh, H. P. Bhattacharyya, M. Sarma, S. Käser, K. Töpfer, L. I. Vazquez-Salazar, E. D. Boittier, M. Meuwly, G. Mandelli, C. Lanzi, R. Conte, M. Ceotto, F. Dietrich, V. Cisternas, R. Gnanasekaran, M. Hippler, M. Jarraya, M. Hochlaf, N. Viswanathan, T. Nevolianis, G. Rath, W. A. Kopp, K. Leonhard and R. A. Mata, *Phys. Chem. Chem. Phys.*, 2023, **25**, 22089–22102.

Open Access Article. Published on 02 September 2025. Downloaded on 9/11/2025 8:59:23 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

12 R. Rahrt, B. Hein-Janke, K. N. Amarasinghe, M. Shafique, M. Feldt, L. Guo, J. N. Harvey, R. Pollice, K. Koszinowski and R. A. Mata, *J. Phys. Chem. A*, 2024, **128**, 4663–4673.
13 **https://drugdesigndata.org/** (last accessed 2025-06-24).
14 H. A. Carlson, R. D. Smith, K. L. Damm-Ganamet, J. A. Stuckey, A. Ahmed, M. A. Convery, D. O. Somers, M. Kranz, P. A. Elkins, G. Cui, C. E. Peishoff, M. H. Lambert and J. B. Dunbar, *J. Chem. Inf. Model.*, 2016, **56**, 1063–1077.
15 S. Ackloo, R. Al-awar, R. E. Amaro, C. H. Arrowsmith, H. Azevedo, R. A. Batey, Y. Bengio, U. A. K. Betz, C. G. Bologa, J. D. Chodera, W. D. Cornell, I. Dunham, G. F. Ecker, K. Edfeldt, A. M. Edwards, M. K. Gilson, C. R. Gordijo, G. Hessler, A. Hillisch, A. Hogner, J. J. Irwin, J. M. Hansen, D. Kuhn, A. R. Leach, A. A. Lee, U. Lessel, M. R. Morgan, J. Moult, I. Muegge, T. I. Oprea, B. G. Perry, P. Riley, S. A. L. Rousseaux, K. Singh Saikatendu, V. Santhakumar, M. Schapira, C. Scholten, M. H. Todd, M. Vedadi, A. Volkamer and T. M. Wilson, *Nat. Rev. Chem.*, 2022, **6**, 287–295.
16 **https://cache-challenge.org/** (last accessed 2025-06-24).
17 A. Nicholls, D. L. Mobley, J. P. Guthrie, J. D. Chodera, C. I. Bayly, M. D. Cooper and V. S. Pande, *J. Med. Chem.*, 2008, **51**, 769–779.
18 J. P. Guthrie, *J. Phys. Chem. B*, 2009, **113**, 4501–4507.
19 N. Tielker, L. Eberlein, O. Beckstein, S. Güssregen, B. I. Iorga, S. M. Kast and S. Liu, in Free Energy Methods in Drug Discovery: Current State and Future Directions, *ACS Symposium Series*, ed. K. A. Armacost and D. C. Thompson, 2021, vol. 1397, pp. 67–107.
20 T. D. Bergazin, N. Tielker, Y. Zhang, J. Mao, M. R. Gunner, K. Francisco, C. Ballatore, S. M. Kast and D. L. Mobley, *J. Comput.-Aided Mol. Des.*, 2021, **35**, 771–802.
21 M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C.'t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao and B. Mons, *Sci. Data*, 2016, **3**, 160018.
22 S. R. Wilkinson, M. Aloqalaa, K. Belhajjame, M. R. Crusoe, B. de Paula Kinoshita, L. Gadelha, D. Garijo, O. J. R. Gustafsson, N. Juty, S. Kanwal, F. Z. Khan, J. Köster, K. Peters-von Gehlen, L. Pouchard, R. K. Rannow, S. Soiland-Reyes, N. Soranzo, S. Sufi, Z. Sun, B. Vilne, M. A. Wouters, D. Yuen and C. Goble, *Sci. Data*, 2025, **12**, 328.
23 B. J. Heil, M. M. Hoffman, F. Markowetz, S. Lee, C. S. Greene and S. C. Hicks, *Nat. Methods*, 2021, **18**, 1132–1135.
24 **https://nfdi4chem.de/** (last accessed 2025-04-10).
25 C. Steinbeck, O. Koepler, F. Bach, S. Herres-Pawlis, N. Jung, J. C. Liermann, S. Neumann, M. Razum, C. Baldauf, F. Biedermann, T. W. Bocklitz, F. Boehm, F. Broda, P. Czodrowski, T. Engel, M. G. Hicks, S. M. Kast, C. Kettner, W. Koch, G. Lanza, A. Link, R. A. Mata, W. E. Nagel, A. Porzel, N. Schlörer, T. Schulze, H.-G. Weinig, W. Wenzel, L. A. Wessjohann and S. Wulle, *Res. Ideas Outcomes*, 2020, **6**, e55852.
26 R. I. Allen, K. J. Box, J. E. A. Comer, C. Peake and K. Y. Tam, *J. Pharm. Biomed. Anal.*, 1998, **17**, 699–712.
27 H. Vatheuer, J. Paulus, L. Johannknecht, G. Keller, R. M. Ziora, L. Stelzl and P. Czodrowski, *ChemMedChem*, 2025, e202500244, DOI: **10.1002/cmdc.202500244**.
28 RDKit: Open-source cheminformatics (ver. 2021.09.2), https://doi.org/10.5281/zenodo.5589557, **https://www.rdkit.org/** (last access 2025-04-10).
29 QUACPAC. OpenEye Scientific Software, Santa Fe, NM.
30 D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, M. Magariños, J. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. Radoux, A. Segura-Cabrera, A. Hersey and A. Leach, *Nucleic Acids Res.*, 2018, **47**, D930–D940.
31 T. Sander, J. Freyss, M. von Korff and C. Rufener, *J. Chem. Inf. Model.*, 2015, **55**, 460–473.
32 M. Işık, A. S. Rustenburg, A. Rizzi, M. R. Gunner, D. L. Mobley and J. D. Chodera, *J. Comput.-Aided Mol. Des.*, 2021, **35**, 131–166.
33 D. T. Manallack, *Perspect. Med. Chem.*, 2007, **1**, 1177391X0700100.
34 C. Liao and M. C. Nicklaus, *J. Chem. Inf. Model.*, 2009, **49**, 2801–2812.
35 J. Manchester, G. Walkup, O. Rivin and Z. You, *J. Chem. Inf. Model.*, 2010, **50**, 565–571.
36 G. T. Balogh, A. Tarcsay and G. M. Keserű, *J. Pharm. Biomed. Anal.*, 2012, **67–68**, 63–70.
37 M. Morgenthaler, E. Schweizer, A. Hoffmann-Röder, F. Benini, R. E. Martin, G. Jaeschke, B. Wagner, H. Fischer, S. Bendels, D. Zimmerli, J. Schneider, F. Diederich, M. Kansy and K. Müller, *ChemMedChem*, 2007, **2**, 1100–1115.
38 J. Noroozi and W. R. Smith, *J. Chem. Eng. Data*, 2020, **65**, 1358–1368.
39 A. C. Lee, J.-Y. Yu and G. M. Crippen, *J. Chem. Inf. Model.*, 2008, **48**, 2042–2053.
40 J. J. Klicić, R. A. Friesner, S.-Y. Liu and W. C. Guida, *J. Phys. Chem. A*, 2002, **106**, 1327–1335.
41 Marvin 21.20, Chemaxon (**https://www.chemaxon.com**).
42 N. Tielker, L. Eberlein, S. Güssregen and S. M. Kast, *J. Comput.-Aided Mol. Des.*, 2018, **32**, 1151–1163.
43 N. Tielker, S. Güssregen and S. M. Kast, *J. Comput.-Aided Mol. Des.*, 2021, **35**, 933–941.
44 RDKit: Open-source cheminformatics. (ver. 2018.03.3), https://doi.org/10.5281/zenodo.1314277, **https://www.rdkit.org/**. (last access 2025-04-10).
45 G. Sigalov, A. Fenley and A. Onufriev, *J. Chem. Phys.*, 2006, **124**, 124902.
46 D. A. Case, K. Belfon, I. Y. Ben-Shalom, S. R. Brozell, D. S. Cerutti, T. E. Cheatham, III, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, G. Giambasu, M. K. Gilson, H. Gohlke, A. W. Goetz, R. Harris, S. Izadi, S. A. Izmailov, K. Kasavajhala, A. Kovalenko, R. Krasny, T. Kurtzman, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, V. Man, K. M. Merz,

Open Access Article. Published on 02 September 2025. Downloaded on 9/11/2025 8:59:23 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, A. Onufriev, F. Pan, S. Pantano, R. Qi, D. R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. L. Simmerling, N. R. Skrynnikov, J. Smith, J. Swails, R. C. Walker, J. Wang, L. Wilson, R. M. Wolf, X. Wu, Y. Xiong, Y. Xue, D. M. York and P. A. Kollman, *AMBER 2020*, University of California, San Francisco, 2020.

47 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 1372–1377.

48 P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, *J. Phys. Chem.*, 1994, **98**, 11623–11627.

49 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian 16, Revision C.01*, Gaussian, Inc., Wallingford CT, 2016.

50 T. Kloss, J. Heil and S. M. Kast, *J. Phys. Chem. B*, 2008, **112**, 4337–4343.

51 H. J. C. Berendsen, J. R. Grigera and T. P. Straatsma, *J. Chem. Phys.*, 1987, **91**, 6269–6271.

52 S. M. Kast, J. Heil, S. Güssregen and K. F. Schmidt, *J. Comput.-Aided Mol. Des.*, 2010, **24**, 343–353.

53 S. M. Kast and T. Kloss, *J. Chem. Phys.*, 2008, **129**, 236101.

54 J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J. Comput. Chem.*, 2004, **25**, 1157–1174.

55 **https://gitlab.tu-dortmund.de/kast_ccb/eurosampl/challenge** (last access 2025-04-10).

56 **https://qmbench.net/** (last access 2025-04-10).

57 **https://www.dublincore.org/** (last access 2025-04-10).

58 **https://schema.datacite.org/** (last access 2025-04-10).

59 R. Fraczkiewicz, in *Comprehensive Medicinal Chemistry II*, Elsevier, ed. B. Testa, H. van de Waterbeemd, 2006, pp. 603–626.

60 G. C. Shields and P. G. Seybold,*Computational Approaches for the Prediction of pK_a* values, CRC Press, 2013.

61 O. D. Abarbanel and G. R. Hutchison, *J. Chem. Theory Comput.*, 2024, **20**, 6946–6956.

62 T. Kalliokoski and K. Sinervo, *Mol. Inform.*, 2019, **38**, 1800163.

63 D. Mobley, personal communication, Feb. 2025.