Implicitly perturbed Hamiltonian: a class of versatile and general-purpose molecular representations for machine learning

Amin Alibakhshi^{1,*}, Bernd Hartke¹

Theoretical Chemistry, Institute for Physical Chemistry, Christian-Albrechts-University, Olshausenstr. 40, 24118 Kiel, Germany

Corresponding author: alibakhshi@pctc.uni-kiel.de

Abstract

Unraveling challenging problems by machine learning has recently become a hot topic in many scientific disciplines. For developing rigorous machine-learning models to study problems of interest in molecular sciences, translating molecular structures to quantitative representations as suitable machine-learning inputs plays a central role. Many different molecular representations and the state-of-the-art ones, although efficient in studying numerous molecular features, still are sub-optimal in many challenging cases, as discussed in the context of present research. The main aim of the present study is to introduce the Implicitly Perturbed Hamiltonian (ImPerHam) as a class of versatile representations for more efficient machine learning of challenging problems in molecular sciences. ImPerHam representations are defined as energy attributes of the molecular Hamiltonian, implicitly perturbed by a number of hypothetical or real arbitrary solvents based on continuum solvation models. We demonstrate outstanding performance of machine-learning models based on ImPerHam representations for three diverse and challenging cases of predicting inhibition of the CYP450 enzyme, high precision and transferrable evaluation of conformational energy of molecular systems and accurately reproducing solvation free energies for large benchmark sets.

Introduction

Employing machine learning for studying complicated scientific challenges has recently become a widely accepted approach in science. As a continuously growing field, machine learning has become

a promising tool in studying many diverse areas in molecular sciences, ranging from major topics in life science such as synthetic biology [1], genomics [2], drug discovery [3], and cell biology [4], to major sub-fields of chemistry such as theoretical [5], organic [6], quantum [7,8], polymer [9] and synthetic [10] chemistry.

Despite this diversity in applications, at the heart of all machine-learning approaches in molecular sciences is a reliance on translating molecular structures to quantities understandable by the machine-learning process. These quantities, which are commonly known as molecular representations in computational chemistry and molecular descriptors in cheminformatics, are uniquely defined and are considered as molecular fingerprints. With the molecular representations defined, the main role of machine learning is then to learn the relationship between those representations and the properties of interest.

The earliest example of employing molecular representations for estimating materials properties, to our knowledge, is the group contribution method proposed in the middle of the last century [11]. This method considers a linear or non-linear dependency between molecular properties and the functional groups present in molecules. The success of this simply defined representation in predicting many properties of chemicals for several decades [12-19] has motivated its employment for approximating potential energies of molecular ensembles [20-23], as one of the most extensively studied and, at the same time, most challenging applications of machine learning in theoretical and quantum chemistry [24,25]. Nevertheless, achieving high accuracy in challenging problems like evaluation of conformational energy of molecular systems, typically requires more versatile molecular representations. To that end, a number of efficient representations have been proposed and widely used in recent years, such as atom-centered symmetry functions [26], the bispectrum of the neighbor density [27], smooth overlap of atomic positions [28], and the Coulomb matrix [29]. The popularity of these representations mainly stems from their remarkably fast and straightforward acquisition, typically requiring only elementary calculations on the geometrical data of the molecules.

These more advanced representations, despite being more efficient compared to elementary representations like functional group number and types, still suffer from limited utility in studying many challenging and complicated problems of interest.

For machine-learning evaluation of conformational energies via those representations, despite several decades of progress, achieving quantum-chemical accuracy still has typically remained limited to simple molecular systems consisting of very few atom types [26,30-34]. As the other limitation, the application of machine-learning models based on employing the commonly applied representations is usually restricted only to the specific systems for which they were developed. This makes it necessary to re-train those machine-learning models for new systems, which is a highly demanding task. Finally, the currently defined representations mainly follow *non-physical* functional forms which results in very limited extrapolation capabilities [35].

For machine-learning evaluation of solvation free energy via conventional representations, in addition to the abovementioned limitations, the results obtainable via most of the currently developed models are still beyond chemical accuracy, as recently reviewed by us [36].

All these issues imply the necessity and importance of developing novel and more efficient representations, capable of addressing the above-mentioned shortcomings. For that purpose, appropriate representations have to satisfy a number of prerequisites such as invariance with respect to translation or rotation of the origin of the coordinate system, invariance to permutation of atoms of the same element, and yielding a unique or constant number of quantities independent of number and type of atoms in the system [34]. Ideally, representations should also allow to transfer machine-learning models to different systems, generate a large number of quantities to enhance machine-learning efficiency for diverse complicated properties of interest, and be physically interpretable.

With all these considerations, the main aim of the present study is to introduce the Implicitly Perturbed Hamiltonian (ImPerHam) as a new class of molecular representations to fulfill all the above-mentioned requirements.

The general idea behind ImPerHam originates from our recently developed machine-learning model for evaluation of solvation free energy [37]. In that work, we employed machine learning for a more rigorous integration of the continuum solvation energy components. To that end, in addition to cavity geometrical data, we also employed energy attributes of Hamiltonian perturbed by the implicitly defined solvent as inputs of machine-learning models.

The outstanding efficiency of implicitly perturbed Hamiltonian energy attributes in characterizing the highly challenging case of solvation free energy reported in our recent study [37] motivated us to employ similar quantities computed not only for a single solvent of interest but also for a diverse set of other implicitly defined solvents and use them as molecular representations for other purposes as well.

To clarify the general idea behind the ImPerHam representations, we discuss evaluation of solvation free energy of methane in water as an example. For that purpose, the conventional approaches require computing energy attributes of the methane Hamiltonian implicitly perturbed in water and integrating them based on the conventional approaches in continuum solvation models or via our recently proposed machine-learning model [37]. Based on the new approach, however, we still employ those energy attributes but computed for methane dissolved a number of other arbitrary real or hypothetical solvents and use them as molecular representations for methane.

An obvious advantage of ImPerHam representations stems from the provided possibility of generating an unlimited number of representations via an unlimited choice of solvents. More importantly, the generated representations are actually thermodynamic quantities attributed to energy and free energy. Considering that based on the foundations of thermodynamics, the majority of system properties and thermodynamic quantities can be expressed as functions of energy and free energy, the representations based on these energy functions can be theoretically related to most properties of interest. For this reason, ImPerHam representations can be expected to perform better than the traditional representations mentioned above, which have less direct and less clear-cut relations to the properties of interest. As another direct result, although an unlimited number of solvents can be defined and employed to generate ImPerHam representations, as will be shown later in the present

study, even very few arbitrary but diversely defined solvents can yield efficient machine-learning models for the diverse challenging problems considered in the present study.

The ImPerHam representations can be generated and studied using low-cost quantum mechanical computations or conventional polarizable force fields, which allows fast evaluation of those representations. Nevertheless, the computational cost of generating the ImPerHam representations can be somewhat higher than for other conventional representations discussed earlier. Despite this higher computational cost of computing ImPerHam representations, considering that the developed models based on them are highly transferable and do not require demanding re-training for new systems, we can still consider them an economic and reasonable choice.

In the present study, we demonstrate the efficiency of the machine-learning models based on ImPerHam representations in studying three challenging and extensively required problems in molecular sciences.

The first case study is evaluation of inhibition of a cytochrome P450 (CYP450) enzyme by small molecules. Prediction of CYP450 inhibition by potential drug candidates is a mandatory consideration in drug discovery, due to the crucially important role of these enzymes in metabolizing the majority of the drugs currently found in the market [38]. Inhibition of CYP450 by potential drugs can result in their accumulation in the body and increase the risk of drug-drug interactions [39]. Mibefradil and Cerivastatin are two examples of commercial drugs which have previously retracted from the market for this reason [40,41]. On the other hand, targeted inhibition of specific cytochromes has been proposed as an effective treatment strategy, as has been demonstrated for metastatic prostate tumors [42]. In the present study, we develop machine-learning models to predict inhibition of CYP450 1A2 as one of the important CYP450 members present in liver. The structure of human microsomal CYP450 1A2 obtained by Sansen et. al. [43] is depicted in figure 1.

Fig. 1 Structure of human microsomal CYP450 1A2 enzyme.

In addition to inhibition of a CYP450 enzyme by machine-learning, we also investigate the performance of ImPerHam representations in machine-learning evaluation of solvation free energy as well as conformational energy, benchmarked for large datasets. These case studies are not only

among the most challenging problems in computational chemistry, but also are extensively required in a wide range of scientific fields which makes them ideal for benchmarking the new representations.

Computational details

Benchmark sets

To be able to make a reliable evaluation of the newly proposed representations, we employed large benchmark sets commonly applied for each studied application.

Accordingly, for studying the inhibition of the CYP450 enzyme we exploited the benchmark set provided by Novotarskyi et. al. for comparing different machine-learning models proposed for the same purpose [44]. We used the same compounds considered by them as training and test datasets to train and validate our model, which included 3745 and 3740 samples, respectively. Note that this benchmark data re-use implies that we have not done any molecule-protein docking calculations ourselves, but instead employed machine learning based on our ImPerHam representations to reproduce the inhibitor/non-inhibitor classification contained in this benchmark set.

To benchmark our newly proposed representations in reproducing high-quality molecular energies, we exploited CCSD(T) reference energies computed for multiple configurations of 3525 dimers provided in the DES370K database [45]. To reduce the computational cost of model development, we randomly selected roughly half of the available dimers a priori, which yielded a dataset containing 1155 neutral, 390 positively charged and 177 negatively charged dimers. From the available conformers of each one of the selected dimers, five were selected for developing the machine-learning models, linearly distributed between the highest and lowest energy conformers.

For studying the predictability of solvation free energies, the efficiency of the developed machine-learning models were benchmark using the Minnesota solvation database [46] containing 2493 reference solvation free energy data for binary mixtures of 435 solutes and 91 solvents.

The accuracy of the developed models is reported as mean unsigned error (MUE) in kcal/mol, defined as:

$$MUE = \frac{1}{N} \sum \left(|y_i^{exp} - y_i^{pred}| \right). \quad (1)$$

Generating representations

We generated ImPerHam representations by perturbing the Hamiltonian via the analytical linearized Poisson-Boltzmann (ALPB) solvation model [47], as implemented in Grimme’s Semiempirical Extended Tight-Binding Program Package [48] (xTB). The quantum-mechanical computations were carried out with the GFN2-xTB semiempirical tight-binding method [49], for a fast and computationally inexpensive acquisition and reasonable estimation of the studied energy attributes.

Table 1 List of energy components of the implicitly perturbed Hamiltonian employed as molecular representations.

1	Total energy
2	Total free energy without cavity and hydrogen bonding contributions
3	HOMO-LUMO gap
4	HOMO orbital eigenvalue
5	LUMO orbital eigenvalue
6	SCC energy
7	Gibbs free energy (total)
8	Gibbs free energy (electric)
9	Gibbs free energy (cavitation)
10	Gibbs free energy (hydrogen bond)
11	Free energy shift for infinite dilution
12	Gradient norm
13	Isotropic electrostatic energy
14	Anisotropic exchange correlation
15	Isotropic exchange correlation
16	Dispersion energy
17	Repulsion energy
18	Atomization energy

For each molecule in the vacuum state as well as in implicitly defined acetone, acetonitrile, benzene, hexane, water, ether, ethyl acetate, 1-octanol and phenol solvents, the Hamiltonian energy attributes listed in table 1, were computed with original parameterizations as provided in xTB [48]. This constituted the initial pool of representations.

These solvents were selected from a wider choice of available solvents parameterized in xTB, with the intention to span a diverse range of dielectric constants. Nevertheless, any real or hypothetical solvent with an arbitrary or actual dielectric constant can also be used for this purpose, resulting in a wide range of relevant representations and higher flexibility of the machine-learning models to evaluate different quantities. This limited number of solvents was considered only to present the approach and efficiency of the proposed representations, and as proof of concept.

For the computed Hamiltonians, the energy attributes reported in table 1 were considered as potentially relevant representations. Nevertheless, many other relevant energy attributes can also be defined and employed for this purpose.

For the machine-learning models developed to evaluate conformational energies via ImPerHam representations, among the 5 selected configurations of each individual dimer, the conformation with

lowest energy was considered as the reference and its energy was set to zero. The energies of the other conformations of the same dimer were also corrected accordingly. This merely is the standard practice in most computational chemistry applications, to get away from total energies (that are irrelevant in most cases, and also frequently method-dependent). Here it focuses the machine-learning approach on the actually relevant relative energies between conformers.

By the above-mentioned treatment of the CCSD(T) conformational energies, the computed representations also were corrected the same way, i.e. the deviations of computed representations for each conformer with respect to the one with the lowest total energy were considered as the inputs to the machine-learning models.

For evaluation of solvation free energies, considering that the reference data were defined as the difference between the free energies of two liquid and gas states, we did not employ the same modification of energy and representations. However, in addition to the computed ImPerHam representations, we also considered the dielectric constant of solvents in each solute-solvent mixture as an additional model input, to characterize the specific solvent under study, similar to conventional continuum solvation models and our recently developed machine-learning model [36].

To compare the accuracy of solvation free energies obtained by machine learning with those obtained via conventional solvation models, we also computed the solvation free energies via the widely-accepted SMD [50], PCM [51], and CPCM [52] continuum solvation models at the B3LYP/6-311+g(2d,p) level of theory in Gaussian 16 [53]. To that end, the geometry of the solutes was first optimized and then used for evaluation of solvation free energy.

Developing machine-learning models

For a more efficient and tractable training of machine-learning models, we employed variable selection as a commonly used approach in developing machine-learning models. To that end, the initial pool of computed representations was screened employing the MRMR variable selection algorithm [54] to yield different sets of input variables containing 4 to 45 representations.

After computation and screening of the required representations based on the above recipe, the machine-learning models taking those representations were set up to study the applications of interest.

Considering that the inhibition of CYP450 enzyme by potential drug candidates is a classification problem, i.e., the role of machine learning is only to classify the studied molecules as inhibitor or non-inhibitor, we employed the Support-Vector-Machine (SVM) classification [55] as a rigorous machine-learning method developed for this purpose. To that end and to be able to make a reliable comparison to the large number of machine-learning models studied by Novotarskyi et. al. [44] for the same purpose, we used the same training and test samples employed in that work.

For evaluation of conformational energy and solvation free energy, we employed artificial neural networks as a highly efficient machine-learning tool to map the dependency between required quantities and the proposed representations. To that end, training of the ANN models was carried out based on the guidelines provided in our recent study [56].

Accordingly, we assigned 60% of the dataset for training, 15% for validation and 25% for testing the machine-learning models. We only studied small neural networks with one hidden layer and a very limited number of neurons. The upper limit of selected number of neurons was assigned to allow having roughly 10 training samples or more per ANN constant. In many neural network models, much lower ratios of training samples to ANN constants are commonly employed, which can result in much higher accuracies for the obtained results. However, this may lead to reducing the extrapolation capability of the neural networks [56]. Therefore, we only considered a limited number of neurons with the upper limit discussed above.

For the machine-learning models developed to evaluate conformational energies, training, validation, and test sets were assigned based on the individual dimer types and not individual conformers, to avoid unreliable high-accuracy results due to interpolation. In other words, if a dimer was assigned to a training, validation or test set, all its conformers were also assigned to the same set.

Results and discussion

Evaluation of CYP450 inhibition

By training SVM models via 3745 training samples provided by Novotarskyi et. al. [44], the predictability of inhibitor activity of 3740 test set molecules was evaluated using the developed machine-learning models. According to the results, the best prediction of inhibitor activity for the training samples was observed for a SVM model that employed 16 representations as model inputs. Via this model, the inhibitor activity of test set compounds could be predicted with 79.6% accuracy which is very close to the results reported for the same datasets by the most accurate models studied by Novotarskyi et. al. [44]. However, the models studied by Novotarskyi et. al. employed a much greater number of representations, ranging from several hundred to roughly five thousand representations.

Considering that our employed representations are generated only for a few solvents, it is expected that for further extensions of the number of solvents, more accurate evaluation of CYP450 inhibition can be achievable.

Machine-learning evaluation of conformational energies via ImPerHam representations

By employing the GFN2-xTB semiempirical tight-binding method to evaluate relative energies of different conformers, we initially obtained an MUE of 22.164 kcal/mol compared to the CCSD(T) reference energies.

Despite this large MUE of the originally evaluated energies by GFN2-xTB computations, a remarkable improvement in accuracy of predicted energies was achieved via the developed machine-learning models based on different combinations of ImPerHam representations. According to the

Fig. 2 Comparison of reference CCSD(T) energies for dimer conformers with energies evaluated by machine learning based on ImPerHam representations and computed by GFN2-xTB, respectively. Employing machine learning and ImPerHam representations improves the MUE from 22.164 kcal/mol to 1.400 kcal/mol, as a single-number measure of the overall deviation from the ideal line indicated in these two graphs.

results, the studied machine-learning models could yield an MUE of 1.400 kcal/mol via a neural network model taking 38 ImPerHam representations as model inputs and 16 neurons in the hidden layer. These results show an improvement in originally computed energies based on the GFN2-xTB tight-binding method by one order of magnitude. A comparison of the reference CCSD(T) configuration energies and the machine-learned and GFN2-xTB computed ones are depicted in figure 2. As further illustration, for the three dimers with the greatest energy variations between their conformers, the CCSD(T) energies are compared with machine learning and GFN2-xTB evaluated energies in different conformers in figure 3.

Table 2 Comparison of the results of the new method with other models for solvation free energy prediction.

Method	Source	Nr. Samples	Nr. Solvents	Nr. Solutes	Deviation measure	Deviation (kcal/mol)
Machine learning	Present study	2493	91	435	MUE RMSE	0.545 0.739
COSMO-RS	Klamt and Diedenhofen [57]	2346	91	318	MUE RMSE	0.42145 0.69644
SM12	Marenich et al. [58]	2403	91	352	MUE	0.5457-0.6717
DCOSMO-RS	Klamt and Diedenhofen [57]	2346	91	318	MUE RMSE	0.6584 0.99724
Feature Functional Theory	Wang et. al. [59]	668	1 (water)	668	RMSE	1.05
kernel-based machine learning	Rauer and Bereau [60]	355	1 (water)	355	MUE	1.06
atoms-in-molecules neural network	Zubatyuk et.al. [61]	—	—	414	MUE	1.1
Structure-Property Relationship	Hutchinson and Kobayashi [62]	—	1 (water)	—	RMSE	1.65
SMD	Present study	2493	91	435	MUE RMSE	0.78623 1.1633
CPCM	Present study	2493	91	435	MUE RMSE	2.6942 3.1733
PCM	Present study	2493	91	435	MUE RMSE	2.9054 3.3948

Machine-learning evaluation of solvation free energies via ImPerHam representations

Among the studied machine-learning models for evaluating the solvation free energies, the best result was obtained for a neural network with only 5 neurons in the hidden layer and 22 ImPerHam representations and solvent dielectric constants as model inputs. For this model we obtained an MUE of 0.545 kcal/mol. Compared to the accuracy of the original ALPB, for which an MUE of 1.4 kcal/mol has been reported [47], our results show a substantial improvement by almost a factor of three. Further comparison of the obtained results with conventionally accepted continuum solvation models are reported in table 2. According to these results, the machine-learning evaluation of the solvation free energy via ImPerHam representations is significantly more accurate than SMD, PCM and CPCM as the most extensively applied continuum solvation models, though obtained for a computational cost reduced by several orders of magnitude (a few seconds compared to several minutes to hours on a normal desktop PC). The distribution of solvation free energies predicted via the newly proposed machine-learning method and by the SMD solvation model, in comparison to experimentally determined data, is depicted in figure 4.

Among the other solvation models reported in table 2, only for the commercially available COSMO-RS solvation model higher accuracies have been reported. However, it should be noted that the results reported here are presented only as a proof of concept. By trying more extensive sets of solvents, larger neural networks and more demanding training of neural network models to search for

Fig. 3 Comparison of CCSD(T) energies with machine learning (ML) and GFN2-xTB evaluated energies for different conformers of three studied dimers with highest variation of energy between conformers. In the illustrated configurations, atoms of C, O, H, P, and Mg are presented with gray, red, white, orange, and green, respectively.

Fig. 4 Comparison of solvation free energies predicted via ML and SMD methods and experimental data.

the global minimum in the evaluated MUE, more accurate results by the machine-learning models are expected to be achievable.

Analysis of the performance of studied solvents and representations

As discussed in the methods section, we considered the gas medium as well as 9 solvents, which provided a diverse range of dielectric constants and therefore different perturbations of the molecular Hamiltonians. The resulting energy attributes were considered as inputs for machine-learning models. The studied representations introduced in table 1 were also a very limited subset of potential energy attributes that could be defined and used as molecular representations. By analysis of those developed models that were among the top 10% of all studied models in terms of accuracy, we investigated the performance of individual solvents and representations.

The percentage of presence of each one of the studied mediums in the selected models with highest accuracies is depicted in figure 5.

As can be seen in figure 5, the most effective perturbing mediums are not the same in different case studies. For example, while the energy attributes perturbed by benzene are available in 96% and 92% of the most effective machine-learning models developed for evaluating molecular energies and CYP450 inhibition, respectively, they are present in only 30% of the models which are effective in estimation of solvation free energy. Additionally, the unperturbed Hamiltonian energy attributes as well as those obtained via perturbed Hamiltonian by ethyl acetate, water and octane, are the most efficient ones in all considered case studies. The other interesting observation here is that all the studied solvents are employed in at least one third of the models.

Table 3 Percentage of presence of studied energy attributes in selected models in difference applications.

Representation ID	Conformational energy	Solvation free energy	CYP450 inhibition
1	87.5	15.38	25.45
2	87.5	0	0
3	16.67	61.57	71.82
4	79.17	84.62	71.82
5	79.17	38.46	100
6	75	46.15	0
7	93.75	100	57.27
8	37.5	15.38	100
9	100	100	100
10	45.83	92.31	100
11	16.67	0	0
12	100	84.62	44.55
13	100	15.38	46.36
14	79.17	69.23	71.82
15	16.67	15.38	97.27
16	37.5	92.31	44.55
17	8.33	46.15	13.64
18	79.17	100	0

Fig. 5 Percentage of presence of studied mediums in selected models for different considered applications.

The percentage of presence of considered energy attributes listed in table 1 in the selected models is reported in table 3. Similar to the studied solvents, here also these results show the significant diversity in efficiency of various energy attributes depending on the application. For example, while the representation with id equal to 2 is present in 87.5% of the accurate models developed for evaluation of molecular energies, it has not been employed in any of the selected models for evaluation of solvation free energy or inhibition of CYP450. Similarly, in all case studies we can find solvents or representations which are present in all of the selected models. Therefore, studying a wider range of potential solvents or representations gives both a high flexibility of ImPerHam representations in studying various problems and allows for achieving higher accuracies.

As can be seen in figure 5, accurate prediction of solvation free energy by machine-learning models can be achieved for significantly lower number of solvents. The potential reason for that can be attributed to the fact that solvation free energy mainly depends on the dielectric constant of the solvent and the geometrical shape of the solute molecules [63]. While the former is already present in all machine-learning models as model input, as discussed in the methods section, the latter remains the same for different solvents. For the exact same reason, all of the machine-learning models that yield the best results for evaluation of solvation free energy employ total Gibbs free energy and cavitation Gibbs free energy as model inputs, as can be seen in table 3.

On the other hand, for prediction of conformational energy, the potential energy attributes, such as isotropic electrostatic energy and the norm of the gradient vector, which is a measure of energy difference between the molecular structure under study and the most stable conformation, are obviously the most relevant representations to the total conformational energy. Consequently, these energy attributes are present in 100% of the machine-learning models that yield the highest accuracies

for evaluation of conformational energies. Additionally, unlike the geometrical shapes required by evaluation of solvation free energy, such energy attributes vary more significantly in different solvents. As a result, the highest employment of all studied solvents also is observed for this application.

For inhibition of the CYP450, as it is implied from the physics of the problem, the free energy of solvation of drug candidates in different solvents can determine the free energy of docking the drugs to the active site of the enzyme [64,65]. To that end, the most widely considered solvents are octanol and water, commonly employed as octanol-water partition coefficient [66-70]. Very interestingly, these two solvents are also present in all of the developed machine-learning models with highest efficiency in evaluating the CYP450 inhibition. For the same reason, the energy attributes which are most clearly related to the solvation free energy, such as electric, cavitation and hydrogen bonding components of the solvation free energy, are present in 100% of the machine-learning models developed for this specific application. From that perspective and considering that these energy attributes construct the total solvation free energy, the representation encoding the total solvation free energy becomes redundant and is not employed as extensively as its constituting components.

Conclusion

In the present study, we introduced ImPerHam as highly versatile representations for describing molecular systems and developing advanced machine-learning models. The ImPerHam representations are various energy attributes which are computed via molecular Hamiltonians in vacuum as well as in implicitly defined solvents. We demonstrated high efficiency and accuracy of machine-learning models based on ImPerHam representations in three diverse applications of predicting CYP450 inhibition by candidate molecules, evaluating the conformational energies in multi-atomic systems and solvation free energies of diverse solute-solvent mixtures. Our results have shown the capability of ImPerHam representations in developing transferable machine-learning models applicable to diverse molecular systems.

References:

1. Faulon, J.-L.; Faure, L., In silico, in vitro, and in vivo machine learning in synthetic biology and metabolic engineering. *Current Opinion in Chemical Biology* 2021, 65, 85-92.
2. Liu, J.; Li, J.; Wang, H.; Yan, J., Application of deep learning in genomics. *Science China Life Sciences* 2020, 1-19.
3. Lavecchia, A., Machine-learning approaches in drug discovery: methods and applications. *Drug discovery today* 2015, 20 (3), 318-331.

4. Sommer, C.; Gerlich, D. W., Machine learning in cell biology—teaching computers to recognize phenotypes. *Journal of cell science* 2013, 126 (24), 5529-5539.
5. Berka, K.; Srsen, S.; Slavicek, P., Is Machine Learning the Future of Theoretical Chemistry? *CHEMICKE LISTY* 2018, 112 (10), 640-647.
6. Liu, Y.; Yang, Q.; Li, Y.; Zhang, L.; Luo, S., Application of Machine Learning in Organic Chemistry. *CHINESE JOURNAL OF ORGANIC CHEMISTRY* 2020, 40 (11), 3812-3827.
7. Dral, P. O., Quantum chemistry in the age of machine learning. *The journal of physical chemistry letters* 2020, 11 (6), 2336-2347.
8. Schütt, K.; Gastegger, M.; Tkatchenko, A.; Müller, K.-R.; Maurer, R. J., Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nature communications* 2019, 10 (1), 1-10.
9. Gormley, A. J.; Webb, M. A., Machine learning in combinatorial polymer chemistry. *Nature Reviews Materials* 2021, 1-3.
10. Pflüger, P. M.; Glorius, F., Molecular machine learning: the future of synthetic chemistry? *Angewandte Chemie International Edition* 2020, 59 (43), 18860-18865.
11. Gamson, W.; Watson, K. In *National Petroleum News* 36, Tech. Sect, 1944; p 1944.
12. Joback, K. G.; Reid, R. C., Estimation of pure-component properties from group-contributions. *Chemical Engineering Communications* 1987, 57 (1-6), 233-243.
13. Fredenslund, A., Vapor-liquid equilibria using UNIFAC: a group-contribution method. Elsevier: 2012.
14. Alibakhshi, A.; Mirshahvalad, H.; Alibakhshi, S., A modified group contribution method for accurate prediction of flash points of pure organic compounds. *Industrial & Engineering Chemistry Research* 2015, 54 (44), 11230-11235.
15. He, T.; Li, S.; Chi, Y.; Zhang, H.-B.; Wang, Z.; Yang, B.; He, X.; You, X., An adaptive distance-based group contribution method for thermodynamic property prediction. *Physical Chemistry Chemical Physics* 2016, 18 (34), 23822-23830.
16. Alibakhshi, A.; Mirshahvalad, H.; Alibakhshi, S., Prediction of flash points of pure organic compounds: Evaluation of the DIPPR database. *Process Safety and Environmental Protection* 2017, 105, 127-133.
17. Li, R.; Herreros, J. M.; Tsolakis, A.; Yang, W., Machine learning regression based group contribution method for cetane and octane numbers prediction of pure fuel compounds and mixtures. *Fuel* 2020, 280, 118589.
18. Kibler, R.; Mohrhardt, B.; Zhang, M.; Breitner, L.; Howe, K. J.; Minakata, D., Group Contribution Method to Predict the Mass Transfer Coefficients of Organics through Various RO Membranes. *Environmental science & technology* 2020, 54 (8), 5167-5177.
19. Graziano, B.; Burkardt, P.; Neumann, M.; Pitsch, H.; Pischinger, S., Development of a Modified Joback–Reid Group Contribution Method to Predict the Sooting Tendency of Oxygenated Fuels. *Energy & Fuels* 2021.
20. Clark, J. A.; Santiso, E. E., SAFT- γ -Mie Cross-Interaction Parameters from Density Functional Theory-Predicted Multipoles of Molecular Fragments for Carbon Dioxide, Benzene, Alkanes, and Water. *The Journal of Physical Chemistry B* 2021, 125 (15), 3867-3882.
21. Fayaz-Torshizi, M.; Müller, E. A., Coarse-grained molecular dynamics study of the self-assembly of polyphilic bolaamphiphiles using the SAFT- γ Mie force field. *Molecular Systems Design & Engineering* 2021.

-
22. Lobanova, O.; Mejia, A.; Jackson, G.; Mueller, E. A., SAFT- γ force field for the simulation of molecular fluids 6: Binary and ternary mixtures comprising water, carbon dioxide, and n-alkanes. *The Journal of Chemical Thermodynamics* 2016, 93, 320-336.
 23. Avendano, C.; Lafitte, T.; Galindo, A.; Adjiman, C. S.; Jackson, G.; Mueller, E. A., SAFT- γ force field for the simulation of molecular fluids. 1. A single-site coarse grained model of carbon dioxide. *The Journal of Physical Chemistry B* 2011, 115 (38), 11154-11169.
 24. Behler, J., Perspective: Machine learning potentials for atomistic simulations. *The Journal of chemical physics* 2016, 145 (17), 170901.
 25. Rupp, M., Machine learning for quantum mechanics in a nutshell. *International Journal of Quantum Chemistry* 2015, 115 (16), 1058-1073.
 26. Behler, J.; Parrinello, M., Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters* 2007, 98 (14), 146401.
 27. Bartok, A. P.; Payne, M. C.; Kondor, R.; Csányi, G., Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical review letters* 2010, 104 (13), 136403.
 28. Bartók, A. P.; Kondor, R.; Csányi, G., On representing chemical environments. *Physical Review B* 2013, 87 (18), 184115.
 29. Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A., Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters* 2012, 108 (5), 058301.
 30. Unruh, D.; Meidanshahi, R. V.; Goodnick, S. M.; Csányi, G.; Zimányi, G. T., Training a Machine-Learning Driven Gaussian Approximation Potential for Si-H Interactions. *arXiv preprint arXiv:2106.02946* 2021.
 31. Liu, Y.-B.; Yang, J.-Y.; Xin, G.-M.; Liu, L.-H.; Csányi, G.; Cao, B.-Y., Machine learning interatomic potential developed for molecular simulations on thermal properties of β -Ga₂O₃. *The Journal of Chemical Physics* 2020, 153 (14), 144501.
 32. Rowe, P.; Deringer, V. L.; Gasparotto, P.; Csányi, G.; Michaelides, A., An accurate and transferable machine learning potential for carbon. *The Journal of Chemical Physics* 2020, 153 (3), 034702.
 33. Davidson, E.; Daff, T.; Csany, G.; Finnis, M., Grand canonical approach to modeling hydrogen trapping at vacancies in α -Fe. *Physical Review Materials* 2020, 4 (6), 063804.
 34. Behler, J., Constructing high-dimensional neural network potentials: A tutorial review. *International Journal of Quantum Chemistry* 2015, 115 (16), 1032-1050.
 35. Behler, J., Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Physical Chemistry Chemical Physics* 2011, 13 (40), 17930-17955.
 36. Amin, A.; Bernd, H., Improved Prediction of Solvation Free Energies by Machine-Learning Polarizable Continuum Solvation Model. 2021.
 37. Alibakhshi, A.; Hartke, B., Improved prediction of solvation free energies by machine-learning polarizable continuum solvation model. 2021.
 38. Wu, Z.; Lei, T.; Shen, C.; Wang, Z.; Cao, D.; Hou, T., ADMET evaluation in drug discovery. 19. Reliable prediction of human cytochrome P450 inhibition using artificial intelligence approaches. *Journal of chemical information and modeling* 2019, 59 (11), 4587-4601.
 39. Miners, J. O.; Mackenzie, P. I.; Knights, K. M., The prediction of drug-glucuronidation parameters in humans: UDP-glucuronosyltransferase enzyme-selective substrate and inhibitor probes for reaction phenotyping and in vitro-in vivo extrapolation of drug clearance and drug-drug interaction potential. *Drug metabolism reviews* 2010, 42 (1), 196-208.
 40. Lasser, K. E.; Allen, P. D.; Woolhandler, S. J.; Himmelstein, D. U.; Wolfe, S. M.; Bor, D. H., Tim-

ing of new black box warnings and withdrawals for prescription medications. *Jama* 2002, 287 (17), 2215-2220.

41. Backman, J. T.; Wang, J. S.; Wen, X.; Kivistö, K. T.; Neuvonen, P. J., Mibefradil but not isradipine substantially elevates the plasma concentrations of the CYP3A4 substrate triazolam. *Clinical Pharmacology & Therapeutics* 1999, 66 (4), 401-407.

42. Porubek, D., CYP17A1: a biochemistry, chemistry, and clinical review. *Current topics in medicinal chemistry* 2013, 13 (12), 1364-1384.

43. Sansen, S.; Yano, J. K.; Reynald, R. L.; Schoch, G. A.; Griffin, K. J.; Stout, C. D.; Johnson, E. F., Adaptations for the Oxidation of Polycyclic Aromatic Hydrocarbons Exhibited by the Structure of Human P450 1A2. *Journal of Biological Chemistry* 2007, 282 (19), 14348-14355.

44. Novotarskyi, S.; Sushko, I.; Körner, R.; Pandey, A. K.; Tetko, I. V., A comparison of different QSAR approaches to modeling CYP450 1A2 inhibition. *Journal of chemical information and modeling* 2011, 51 (6), 1271-1280.

45. Donchev, A. G.; Taube, A. G.; Decolvenaere, E.; Hargus, C.; McGibbon, R. T.; Law, K.-H.; Gregersen, B. A.; Li, J.-L.; Palmo, K.; Siva, K., Quantum chemical benchmark databases of gold-standard dimer interaction energies. *Scientific data* 2021, 8 (1), 1-9.

46. Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G., Minnesota solvation database. *Minnesota Solvation Database version* 2012, 20.

47. Ehlert, S.; Stahn, M.; Spicher, S.; Grimme, S., A robust and efficient implicit solvation model for fast semiempirical methods. 2021.

48. Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S., Extended tight-binding quantum chemistry methods. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 2021, 11 (2), e1493.

49. Bannwarth, C.; Ehlert, S.; Grimme, S., GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of chemical theory and computation* 2019, 15 (3), 1652-1671.

50. Marenich, A. V.; Cramer, C. J.; Truhlar, D. G., Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *The Journal of Physical Chemistry B* 2009, 113 (18), 6378-6396.

51. Mennucci, B.; Cammi, R.; Tomasi, J., Excited states and solvatochromic shifts within a nonequilibrium solvation approach: A new formulation of the integral equation formalism method at the self-consistent field, configuration interaction, and multiconfiguration self-consistent field level. *The Journal of chemical physics* 1998, 109 (7), 2798-2807.

52. Barone, V.; Cossi, M., Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *The Journal of Physical Chemistry A* 1998, 102 (11), 1995-2001.

53. Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Petersson, G.; Nakatsuji, H., *Gaussian 16. Revision A* 2016, 3.

54. Peng, H.; Long, F.; Ding, C., Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence* 2005, 27 (8), 1226-1238.

55. Jakkula, V., Tutorial on support vector machine (svm). *School of EECS, Washington State University* 2006, 37.

-
56. Alibakhshi, A., Strategies to develop robust neural network models: Prediction of flash point as a case study. *Analytica chimica acta* 2018, 1026, 69-76.
57. Klamt, A.; Diedenhofen, M., Calculation of solvation free energies with DCOSMO-RS. *The Journal of Physical Chemistry A* 2015, 119 (21), 5439-5445.
58. Marenich, A. V.; Cramer, C. J.; Truhlar, D. G., Generalized born solvation model SM12. *Journal of Chemical Theory and Computation* 2013, 9 (1), 609-620.
59. Wang, B.; Wang, C.; Wu, K.; Wei, G. W., Breaking the polar-nonpolar division in solvation free energy prediction. *Journal of computational chemistry* 2018, 39 (4), 217-233.
60. Rauer, C.; Bereau, T., Hydration free energies from kernel-based machine learning: Compound-database bias. *The Journal of Chemical Physics* 2020, 153 (1), 014101.
61. Zubatyuk, R.; Smith, J. S.; Leszczynski, J.; Isayev, O., Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Science advances* 2019, 5 (8), eaav6490.
62. Hutchinson, S. T.; Kobayashi, R., Solvent-specific featurization for predicting free energies of solvation through machine learning. *Journal of chemical information and modeling* 2019, 59 (4), 1338-1346.
63. Alibakhshi, A., Thermodynamically effective molecular surfaces for more efficient study of condensed-phase thermodynamics. 2021.
64. Azam, M. A.; Saha, N.; Jupudi, S., An explorative study on *Staphylococcus aureus* MurE inhibitor: induced fit docking, binding free energy calculation, and molecular dynamics. *Journal of Receptors and Signal Transduction* 2019, 39 (1), 45-54.
65. Oliveira, F. G.; Sant'Anna, C. M.; Caffarena, E. R.; Dardenne, L. E.; Barreiro, E. J., Molecular docking study and development of an empirical binding free energy model for phosphodiesterase 4 inhibitors. *Bioorganic & medicinal chemistry* 2006, 14 (17), 6001-6011.
66. Valencia-Islas, N. A.; Arguello, J. J.; Rojas, J. L., Antioxidant and photoprotective metabolites of *Bunodophoron melanocarpum*, a lichen from the Andean páramo. *Pharmaceutical Sciences* 2020, 27 (2), 281-290.
67. Ghamri, M.; Harkati, D.; Belaidi, S.; Boudergua, S.; Said, R. B.; Linguerrri, R.; Chambaud, G.; Hochlaf, M., Carbazole derivatives containing chalcone analogues targeting topoisomerase II inhibition: First principles characterization and QSAR modelling. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 2020, 242, 118724.
68. Stępnik, K.; Kukula-Koch, W., In Silico Studies on Triterpenoid Saponins Permeation through the Blood–Brain Barrier Combined with Postmortem Research on the Brain Tissues of Mice Affected by Astragaloside IV Administration. *International journal of molecular sciences* 2020, 21 (7), 2534.
69. Ventura, F. F.; Mendes, L. F.; Oliveira, A. G.; Bazito, R. C.; Bechara, E. J.; Freire, R. S.; Stevani, C. V., Evaluation of phenolic compound toxicity using a bioluminescent assay with the fungus *Geronomema viridilucens*. *Environmental toxicology and chemistry* 2020, 39 (8), 1558-1565.
70. Gunesch, A. P.; Zapatero-Belinchón, F. J.; Pinkert, L.; Steinmann, E.; Manns, M. P.; Schneider, G.; Pietschmann, T.; Brönstrup, M.; von Hahn, T., Filovirus antiviral activity of cationic amphiphilic drugs is associated with lipophilicity and ability to induce phospholipidosis. *Antimicrobial agents and chemotherapy* 2020, 64 (8), e00143-20.