MASARYKOVA
UNIVERZITA

MASARYK UNIVERSITY
Faculty of Science

Ph.D. Thesis

Partial Atomic Charges and Their Chemoinformatics Application

Brno 2021

Stanislav Geidl

To my grandma.

Acknowledgements

Declaration

I hereby declare that this disertation thesis is my original authorial work, which I have worked on alone. All sources, references and literature used or excerpted during the elaboration of this work are properly cited and listed in a complete reference with regard to the source.

Brno, xxth xxxxxxx 2021 ..... Stanislav Geidl

Contents

Part I

Introduction

1

Introduction

In recent years, a vast amount of data about various types of molecules became available. For example, we can obtain the complete human genome of a selected individual in a few days, and about 150 thousand biomacromolecular structures have been determined and published (Protein Data Bank [1]). Furthermore, more than 100 million various small molecules are described in freely accessible databases (e.g., Pubchem [2], ZINC [3], ChEMBL [4]). This richness of data caused the formation of novel modern life-science research fields focused on the utilization of this data. The best-known modern life sciences are bioinformatics, structural bioinformatics, systems biology, genomics, proteomics, and also chemoinformatics. These current research specializations have provided many key results in basic and applied research (e.g. [5–11]).

One fascinating and beneficial field utilizing and processing newly available data about small molecules (i.e., drug-like compounds) is chemoinformatics. This discipline offers methodologies for comparing molecular similarity, molecular database search, virtual screening, and the prediction of molecules' properties and activities. This prediction is based on the idea that molecular structures' similarity has a consequence – a similarity in molecular properties. In chemoinformatics, the structure is first described using mathematical characteristics (so-called descriptors) – numbers containing 3D (or 2D or 1D) structure information and applicable as inputs of mathematical models. Then, these models are constructed based on a relation between descriptors and known values of the property or the activity. Such models are called Quantitative Structure-Property Relationship (QSPR) models or Quantitative Structure-Activity Relationship (QSAR) models.

A property, which is strongly required and is therefore often a target of chemoinformatics prediction models is the acid dissociation constant, K_a, and its negative logarithm pK_a. Those pK_a values are of interest in chemical, biological, environmental, and pharmaceutical research [12–14]. pK_a values have found applications in many areas, such as evaluating and optimizing drug candidate molecules, pharmacokinetics, ADME profiling, understanding protein-ligand interactions, etc. Moreover, the critical physicochemical properties such as permeability, lipophilicity, solubility, etc., are pK_a dependent. Unfortunately, experimental pK_a values are available only for a limited set of molecules. In addition to that, obtaining experimental pK_a values for newly designed molecules is very time-consuming because they must be synthesized first. Chemoinformatics approaches for pK_a prediction are therefore currently intensively examined.

For this reason, I also focused on the chemoinformatics way of pK_a prediction in my work. Very promising descriptors for pK_a prediction are partial atomic charges [] because they hold

information about the distribution of electron density within the molecule. Specifically, electron densities on atoms close to the dissociating hydrogen provide a clue about its dissociation ability. The most common and accurate method for calculating partial atomic charges is an application of quantum mechanics (QM). QM calculation can be performed via various approaches, introducing different approximation levels (i.e., approximating a wave function by different sets of mathematical equations, which are called basis sets). QM outputs electron distribution in orbitals and this distribution can be divided into individual atoms using several charge calculation schemes (e.g., MPA, NPA, AIM, Hirshfeld, MK, etc.). Therefore, the correlation between pK_a and relevant atomic charges calculated by different QM approaches has been analyzed []. I also focused on this file in my bachelor thesis, developed a workflow for calculation of pK_a using QM partial atomic charges and examined, which types of QM are the most suitable [].

QM charges are accurate, but their calculation is very time-consuming. A faster Alternative to QM charges is empirical charge calculation approaches. Furthermore, if we would like to apply chemoinformatics pK_a prediction models practically – for example, in pre-screening large sets of drug candidates – we need a fast approach. Therefore, in my master thesis, I developed a pK_a prediction workflow based on charges (including Electronegativity Equalization Method).

However, several pieces of the puzzle were still missing. For example, the developed pK_a prediction workflows [] were strongly dependent on 3D structure source, and also, the quality of available EEM charges was low.

Therefore, my dissertation's goal was to develop a workflow that predicts pK_a for molecules not synthesized yet and without available experimental 3D structures.

Specifically, the thesis examined how to improve the process of pK_a prediction via providing suitable inputs. First, the influence of 3D structure source on pK_a prediction accuracy was analyzed. Afterward, the work focused on obtaining high-quality partial atomic charges, which served as descriptors for pK_a calculation. In the end, the authors also support the development of methodology and software tools for obtaining these high-quality charges.

The thesis structure is the following: First, an overview of key fields is provided (Part II), i.e. – 3D structure and approaches for its prediction, charge calculation methods, and pK_a prediction approaches. Next, the achieved results, which we published in three research papers, are briefly described (Part III), and full-texts of the respective published papers are attached in Appendix: Main papers. During the elaboration of this thesis, I was also involved in other projects. Most of them were not related to pK_a prediction but tightly connected to the field of chemoinformatics or structural bioinformatics. The outcome of these projects consists of several papers and a book I have co-authored. Their title pages are attached in Appendix: All My Publications.

Part II

Theory and Methods

2

Structure

2.1 Molecular Structure in Computer

The chemical structure is the essential information for chemoinformatics and computational chemistry calculations. We recognize different types of chemical structures according to the complexity of information [15].

The empirical or chemical formula provides information about molecule composition – elements and their count. The structural formula (2D structure) extends this information about topology – bonds between atoms. The three-dimensional structure also provides the conformation of a molecule – the relative placement of atoms in space. We try to provide conformation with the lowest energy representing the most probable conformation of molecule in reality. For some applications, there can even be an assembly of these 3D structures.

In chemoinformatics, two-dimensional structures are often used, but the three-dimensional structure can often bring new information into the *in silico* calculations or models. On the other hand, this 3D structure can be obtained experimentally for a limited number of small molecules. What with other molecules, including those which were not synthesized yet?

2.2 3D Structure Calculation

We apply more computationally efficient methods for 3D structure computation because we use them as input for high-throughput methods. For this reason, many resources were devoted to the development of fast and accurate 3D structure prediction methods [**?**]. These can be classified into the following groups []: rule-based and data-based, fragment-based, numerical methods, and conformational analysis. These rule-based and data-based, fragment-based methods are partially overlying.

2.2.1 Rule-Based and Data-Based Methods

These methods [] use chemical knowledge of geometric and energetic rules known from experiments and theoretical calculations. In these methods, we use rules explicitly to describe, e.g., bond lengths and angles; we use data implicitly to describe, e.g., ring conformation.

2.2.2 Fragment-Based Method

The fragment-based method [] is the incremental method using rules in the first step to fragment a structure into parts. According to the following rules, the parts are assembled by linking fragment templates from a library (database). Predicted structures are created from the most similar and largest fragments in a database as possible.

2.2.3 Numerical Method

These numerical methods [] consist of three methods: molecular mechanics (MM), quantum mechanics, and distance geometry (DG). Distance geometry is a great tool to prepare a reasonable initial structure, which is very close to some low-energy conformation. For this structure, we can use the optimization process from MM or/and QM.

2.2.4 Conformational Analysis

This method [] generates a set of conformations for one molecule using different approaches - genetic algorithms, systematic methods, random techniques, Monte Carlo or MD simulation. The one or more different structures are selected based on criteria such as the number of conformers, minimum RMSD [], only conformations with the lowest MM energy (low-energy conformers).

3

Partial Atomic Charges

3.1 The Concept of Atomic Charges

Atomic charges are a theoretical concept for the quantitative description of electron density around every atom in a molecule. The first basic concept came from early chemistry, where an integer expressed these charges (e.g. -1, +2). Later, they were a real number (partial charge) in organic chemistry and physical chemistry [16]. It is a great approach to explain the mechanism of a lot of chemical reactions. Recently, partial atomic charges also became popular in chemoinformatics, as they proved to be informative descriptors for QSAR and QSPR modeling [17, 18] and for other applications [19–21]; they can be utilized in virtual screening [22, 23] and similarity searches [24, 25]. In reality, we are not able to measure these numbers, only calculate or estimate them. For such reasons, many different approaches for the calculation of partial atomic charges were developed.

3.2 Overview of Charge Calculation Methods

3.2.1 QM Charge Methods

These methods use a wave function as a starting point and then apply subsequent population analysis, charge calculation scheme, or fit to some physical observation [**?**].

Mulliken population analysis (MPA) [26, 27] simply calculates a charge of an atom as a sum of an electron density from its molecular orbitals and a half of an electron density from its bonding orbitals. Natural population analysis (NPA) [28, 29] sophisticatedly improves the MPA method by orthogonalization of specific atoms and after this, NPA performs charge assignment from electron density the same way as in MPA. NPA atomic charges are more stable and independent of the size of basis sets. Other possible population analyses are Löwdin population analysis [30], Hirshfeld population analysis [31].

AIM (atoms-in-molecules) charge calculation scheme is based on the idea that electron density measured by X-ray can help with the calculation of partial charges. Bader and his coworkers [32, 33] defined an atomic volume that is used for charge calculation. Other well-known approaches are electrostatic potential fitting methods (ESP) like CHELPG [34] or MK (Merz-Singh-Kollman) [35] and their extension – RESP methods [**?**].

Cramer and at [36] also developed semiempirical methods – charge model 5 (CM5), which extends Hirshfeld population analysis by empirical parameters to reproduce charge-dependent observables.

3.2.2 Empirical Methods

Empirical approaches use only empirical parameters, and some of these can calculate charges from the 3D structure or only from the topology (2D structure) of a molecule. Therefore, they are distinctly faster than QM approaches.

One of the first empirical methods developed, CHARGE [67], performs a breakdown of the charge transmission by polar atoms into single-bond, double-bond, and triple-bond additive contributions. Other empirical methods have been developed on the electronegativity equalization principle. One group of these empirical approaches are using the Laplacian matrix formalism and the product is a redistribution of electronegativity: Gasteiger-Marsili (PEOE, partial equalization of orbital electronegativity) [37,38], GDAC (geometry-dependent atomic charge) [**?**], KCM (Kirchhoff charge model) [39], DENR (dynamic electronegativity relaxation) [40] or TSEF (topologically symmetric energy function) [40].

The second group of approaches applies the full equalization of orbital electronegativity. For example, this group contains EEM (electronegativity equalization method) [41] and its extensions (ABEEM [**?**], SFKEEM [17]), QEq (charge equilibration) [39], EQEq (extended QEq) [**?**], or SQE (split charge equilibration) [40].

Group of conformationally independent methods (based on the 2D structure) contains CHARGE, Gasteiger-Marsili, KCM, DENR, and TSEF. Group of conformationally dependent – geometrical charges (based on the 3D structure) also consider an influence of conformation and includes the following methods: GDAC, EEM, ABEEM, SFKEEM, QEq, EQeq, and SQE.

A typical representative of the topological method is the Gasteiger-Marsili method, which first assigns charges based on atom types and then iteratively updates atomic charges based on the closest partners. The correction is smaller and smaller in every step until the sixth step when these corrections are too small and atomic charges are final. Empirical parameters for this method were calculated from QM.

On the other hand, the EEM method needs a complete 3D structure and more applicable charges for some of the applications.

3.3 EEM Calculation

EEM (electronegativity equalization method) [41] is one of the most popular empirical charge calculation methods and was developed more than twenty years ago. This method's new parameterizations [D17, D56–D62] and extension [D59, D63, D64] are still under development. An advantage of EEM calculation is that it considers the influence of the molecule's conformation on the atomic charges. For this reason, EEM charges are often used in predictive models as chemoinformatics regressors (descriptors) [D65].

EEM is based on three principles:

The first principle is Sanderson's electronegativity equalization principle. It assumes that the effective electronegativity of each atom in the molecule is equal to the molecular electronegativity:

$$\chi_1 = \chi_2 = \cdots = \chi_x = \bar{\chi} \tag{3.1}$$

where χ_x is the effective electronegativity of the atom x and $\bar{\chi}$ is the molecular electronegativity.

The second principle is the principle of the charge balance. The sum of all charges is equal to the total charge Q:

$$\sum_{x=1} q_x = Q \tag{3.2}$$

where q_x is the charge of the atom x.

And the last principle is the principle of charge-dependent electronegativity. This principle is the definition of atomic electronegativity, and states that the electronegativity of each atom can be expressed as a function of its charge:

$$\chi_i = A_i + B_i \cdot q_i + \kappa \sum_{j=1\, i\neq j}^{N} \frac{q_j}{R_{i,j}} \tag{3.3}$$

where $R_{i,j}$ is the distance between atoms i and j, and the coefficients A_i, B_i and κ are empirical parameters.

These principles can be summed up to a system of equations with N + 1 unknowns (where $q_1, q_2, \ldots, q_N$ and $\bar{\chi}$):

$$\begin{pmatrix} B_1 & \frac{\kappa}{R_{1,2}} & \cdots & \frac{\kappa}{R_{1,N}} & -1 \\ \frac{\kappa}{R_{2,1}} & B_2 & \cdots & \frac{\kappa}{R_{2,N}} & -1 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \frac{\kappa}{R_{N,1}} & \frac{\kappa}{R_{N,2}} & \cdots & B_N & -1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_N \\ \bar{\chi} \end{pmatrix} = \begin{pmatrix} -A_1 \\ -A_2 \\ \vdots \\ -A_N \\ Q \end{pmatrix} \tag{3.4}$$

The first values of parameters A_i and B_i were modifications of experimental hardness and electronegativity [41]. κ was equal to 1. Nowadays, these parameters are calculated from the QM charges [D17, D56–D62]. Therefore, EEM charges were correlated with the QM charge calculated with the same method used for parametrization.

3.4 Quality and Usability of EEM parameters

The quality of EEM parameters describes how the empirical charges computed using these EEM parameters correspond with QM charges used for EEM parameterization. Three main characteristics can describe the quality of EEM parameters – the coefficient of determination [D69, D70] root mean square error (RMSE) [D69, D70] and average absolute error ($\bar{\Delta}$).

The coefficient of determination R^2 is the squared value of the Pearson coefficient (equation 3.5). This value describes the linear correlation rate. Values close to 1 mean that values correlate very well, and values close to 0 mean no correlation.

$$R = \sqrt{\frac{\sum_{x=1}^{N} \left((q_x^{calc} - \overline{q}_x^{calc}) \cdot (q_x^{ref} - \overline{q}_x^{ref}) \right)}{\sum_{x=1}^{N} (q_x^{calc} - \overline{q}_x^{calc})^2 \cdot \sum_{x=1}^{N} (q_x^{ref} - \overline{q}_x^{ref})^2}} \tag{3.5}$$

where q^{ref} is the reference value of charge calculated by QM and q^{calc} is charge value calculated by EEM. $\overline{q}^{ref}, \overline{q}^{calc}$ are the average value of q^{ref}, respectively q^{calc}.

Root mean square error RMSE is the normalized sum of squared error describing the reliability of the model calculated by:

$$\mathrm{RMSE} = \frac{\sum_{x=1}^{N} (q_x^{calc} - q_x^{ref})^2}{N} \tag{3.6}$$

Average absolute error $\overline{\Delta}$ is an averaged difference between corresponding EEM and QM charges in a molecule and is calculated according to an equation:

$$\overline{\Delta} = \frac{\sum_{x=1}^{N} |q_x^{calc} - q_x^{ref}|}{N} \tag{3.7}$$

Their **coverage** describes the applicability of EEM parameters. Coverage is a percentage value describing EEM parameters' ability to calculate charges for individual molecules in a dedicated dataset. *De facto*, this coverage depends on the representation of atom types in EEM parameters.

$$\mathrm{coverage} = \frac{N_{pos}}{N_{tot}} \tag{3.8}$$

where N_{pos} is the number of molecules able calculated by EEM parameters and N_tot is the total number of molecules in a dataset.

3.5 EEM Parametrization

For the parameterization of EEM charges, a lot of different methods have been introduced []. We can summarize it into two groups: one group contains a method that analytically solves equation x – linear regression [] and the second group contains optimization methods [42] such as Accelerated Random Search, Particle Swarm Optimization, and Differential Evolution algorithms. Both of these groups need input – a set of molecules with 3D structures and QM atomic charges. In my work, linear regression and differential evolution were used, and therefore, they are described in more detail below:

The linear regression (LR) method is based on these two equations:

$$A_i + B_i \cdot x = y \tag{3.9}$$

$$\begin{aligned} x &= q_i \\ y &= \chi_i - \kappa \sum_{j=1\, i\neq j}^{N} \frac{q_j}{R_{i,j}} \end{aligned} \tag{3.10}$$

Equations are derived from equations **??** and **??**, which define the EEM method. In the LR method, the dataset of molecules with QM charges can change in every iteration to improve the quality of resulting charges. Quality criterium can be the Pearson correlation coefficient or the coefficient of determination, and the root mean square error or different types of errors. An advantage of the LR method is its straightforwardness and the possibility to optimize κ by another iteration. On the other hand, this method is not possible to make parametrization for some extensions of EEM like SFKEEM and ABEEM.

Differential Evaluation (DE) [**?**] is a heuristics method to focus on finding a global minimum of a function. This method works similar to other optimization methods – iteratively optimize parameters to improve the final solution. Parameters of function are set up randomly, mutated, and evaluated until there is no best solution.

4

Acid Dissociation COnstant Prediction

4.1 Motivation

The acid dissociation constant (pK_a) is a physicochemical property that characterizes the strength of acids. It is one of the essential properties for pharmaceutical, chemical, biological and environmental research or industry. For example, it can be used in the chemoinformatics pipeline for evaluation and optimization of drug candidate [43–45], ADME profiling [46, 47], pharmacokinetics [12], understanding protein-ligand interactions [13, 48].

4.2 Overview of Approaches

Several different approaches for pKa prediction have been developed [48–51].

4.2.1 LFER (Linear Free Energy Relationships) Methods

This is one of the oldest methods [52, 53] for pK_a prediction. This method uses the linear relation of Gibbs energy and pK_a or the logarithm of a reaction rate constant – the Hammett and Taft equations. An advantage of this method is a simple, straightforward, and quick calculation, but on the other hand, we need substituent and reaction parameters. This method is still used in the programs ACD/pKa [**?**], EPIK [**?**], and SPARC [**?**].

4.2.2 Database Methods

These methods [**?**, 54] use a library (database) of molecules with known pKa values. The pK_a value is taken directly from this library, or it is interpolated or triangulated from most similar molecules in this library. Most accurate results are produced only for molecules that are similar to molecules in the database. For this reason, it is essential to have an extensive library.

4.2.3 Ab Inition Quantum Mechanical Calculations

These methods [55, 56] use the fact that the dissociation constant can be calculated from the Gibbs energy of the reaction and from the solvation based on equation 4.2. However, there is no general approach, and every specific calculation configuration needs to be calibrated based on experimental values. The significant disadvantage of these methods is that they are time-consuming. On the other hand, these methods can be very accurate if they use correct calibration parameters. It is only one of the few methods that can be used to extend the training dataset with experimental values or validate some of this experimental value. It means that other methods can be improved by this method. This method is implemented as a module of the Jaguar quantum chemical software package [**?**].

$$\mathrm{p}K_a = -\log_{10} K_a \tag{4.1}$$

$$K_a = e^{\frac{-\Delta G^{\bigcirc}}{RT}} \tag{4.2}$$

4.2.4 QSPR Method

The quantitative structure-property relationship method [] [] uses mainly a linear model to describe the relationship between molecular structure and a property of a molecule, in our case pK_a. In those models, structures are presented by descriptors [] that are numerical expressions of molecular properties. For example, descriptors can be the number of hydrogen atoms, the ratio between carbon atoms and all atoms in the molecule, or solvent accessible surface area.

pK_a correlates well with the polarizability, HOMO energy [**?**], proton-transfer energy [35], partial atomic charges [**?**, 18, 57, 58], the electrostatic potential of the molecule [**?**], etc. Partial atomic charges proved as very promising descriptors [**?**, 18, 57, 58] for pK_a prediction.

Part III

Results

5

Synopsis of the Results

We published a series of articles about pKa prediction [] where we showed that some specific atomic charges correlate with pK_a. Based on this, we were able to build QSPR models for the prediction of this property. We also compared QM, EEM charges, and their models. In this dissertation, I focused only on the last one:

Geidl S, Svobodová Vařeková R, Bendová V, Petrusek L, Ionescu C-M, Jurka Z, Abagyan R, Koča J: **How Does the Methodology of 3D Structure Preparation Influence the Quality of pK_a Prediction?** *J Chem Inf Model* 2015, **55**:1088–1097.

In this article, we utilized different approaches to generate the 3D structures of organic molecules. These structures were used for the building of pK_a prediction models based on charge descriptors. Then we analyzed various influences and relationships and found which methodologies for 3D structure preparation are applicable for pK_a prediction.

We examined not only pKa prediction models employing QM charges but also the models utilizing EEM charges. An application of EEM charges looked very promising. Moreover, EEM charge calculation is significantly faster than QM charge calculation.

However, in parallel, we found one significant limitation of EEM charges – the parameters. It was available a reasonable number of parameter sets, but they had only a low coverage. For example, Bultinc's parameter set [**?**] contains parameters only for these elements: C, F, H, N, O. This fact markedly reduced a dataset, which we were able to use for QM charges.

A lack of parameters disallows the usage of EEM charges in many chemoinformatics applications. For this reason, in our follow-up work, we focused on the development of new and more robust EEM parameters. The first step was to develop a new parameterization tool that was easy to use and extendable. After the prototype, we carefully prepared a new dataset of molecules, and for this dataset, we computed EEM parameters with higher coverage. These new parameters were published in a scientific paper:

Geidl S, Bouchal T, Raček T, Svobodová Vařeková R, Hejret V, Křenek A, Abagyan R, Koča J: **High-quality and universal empirical atomic charges for chemoinformatics applications.** *J Cheminform* 2015, **7**:59.

After some tuning up and extensive research, we released and also published the tool for EEM parametrization – the NEEMP software:

Raček T, Pazúriková J, Svobodová Vařekova R, Geidl S, Křenek A, Falginella FL, Horský V, Hejret V, Koča J: **NEEMP: Software for validation, accurate calculation and fast parameterization of EEM charges.** *J Cheminform* 2016.

Sections **??**, **??**, and **??** contain a summary and Appendix: Main papers full text of all these aforementioned articles. I was also involved in other projects, and the outcome of it is a list of additional articles and book chapters in Appendix: All My Publications.

5.1 How Does the Methodology of 3D Structure Preparation Influence the Quality of pKa Prediction?

From our previous articles [], we know that the prediction of pK_a is possible via QSPR models using partial atomic charges as descriptors. The article's goal was to discover how the methodology of 3D structure preparation influences the quality of pKa prediction. We prepared a dataset containing 60 phenols, 82 carboxylic acids, 48 anilines, and additional testing 53 phenols for these purposes. We took structures from 5 different sources for all these molecules: PubChem, DTP NCI database: Balloon, Frog2, OpenBabel, and RDKit software. We used neutral forms of all the molecules and dissociated forms of phenols and carboxylic acids, and associated forms of anilines. We also optimized these structures with MM (Molecular mechanics) and QM. All combination led to 7220 structures for that we calculated four different QM, one semiempirical QM, four different EEM charges, and Gasteiger-Marselli charges. We created 516 QSPR models for all these molecules and charges. The robustness of these models was tested by cross-validation and QM charges also by an independent test set of phenols.

We confirmed that QM and EEM charge descriptors could be used for pK_a prediction. About half of all models have excellent quality with $R2 >= 0.9$. We also showed that ab initio and semiempirical charges correlate with pK_a and their models are very accurate. In EEM, we had models with a little worse quality, but empirical charges are calculated much faster, and an application in chemoinformatics is much more appropriate. In our models, we were not able to use Gasteiger-Marsili charges to get an adequate quality.

We focused on different types of influence. For classes of the benzene derivates (phenols and anilines) was much easier to obtain high-quality models. Nevertheless, for aliphatic hydrocarbon derivates (carboxylic acid), it was more challenging.

The focus of this article was a comparison of the source of the 3D structure. The influence of input structure for models is essential because the result – the quality of QSPR models – depends on input structures and their quality. For example, structures taken from RDKIT generated only by distance geometry produced fragile models. On the other hand, the 3D structures from the DTP NCI and PubChem databases, formally structures generated by CORINA and Omega, exhibited the best performance for all the tested molecular classes and charge calculation approaches. Structures generated by Frog2 also performed very well. Other 3D structure sources can also be used, but with caution.

We also tried the influence of structure optimization on the quality of QSPR models. In most cases, differences between original structures and optimized structures were slight.

In this article, we summed up the best workflow for the fast and accurate prediction of pKa. This or similar workflow can also be easily applied to other important properties for in silico designed molecules. Flow is about preparing 3D structures by CORINA or Omega (with no further optimization), calculation of EEM charges for these structures, and then the EEM QSPR calculation of pK_a.

5.1.1 My contribution

I prepared the input dataset (by extension of published datasets), performed all the calculations, participated in the analysis of the results, and wrote a part of the manuscript, including all tables and graphics.

5.2 High-quality and universal empirical atomic charges for chemoinformatics applications

The EEM method for charge calculation was published several decades ago. Before our study, there were done some improvements of EEM [], parameterizations of empirical parameters [], and developments of new parameterization methods []. However, there was a problem with the usability of EEM because the available parameter sets had a low coverage in chemical space.

We prepared a set of 4475 distinct small organic, drug-like molecules containing these elements: H, C, N, O, F, P, S, Cl, Br, and I, in different functional groups. This set was created so that each selected atom type is contained in at least 100 molecules. CORINA calculated the 3D structure for all molecules. next step was the calculation of reference QM charges. We selected 6 different approaches: B3LYP/6-311G/MPA, B3LYP/6-311G/NPA, B3LYP/6-311G/AIM, HF/6-311G/MPA, HF/6-311G/NPA, and HF/6-311G/AIM. EEM parameterization was performed for six QM charge calculation approaches, and the whole set of prepared molecules was used.

Our new EEM parameters get very high quality – all Pearson coefficients had a value greater or equal to 0.9. We also showed that the used QM approach did not prove any difference in the quality of parameters – B3LYP and HF produced comparable results. EEM parameters based on NPA and AIM population analysis are slightly better than EEM parameters based on MPA.

We also calculated coverage of our parameters previously published parameters on four big chemoinformatics databases of drug-like molecules — DrugBank, ChEMBL, PubChem, and ZINC. We found out that their coverage is less than 60% for most of the previously published parameters. Our newly produced parameters showed coverage of at least 90% for these databases. Consistency of coverage points out that this problem is not related to a database but concerns a chemical space of drug-like molecules and their atom types.

For evaluation of quality, we selected 657 approved drugs from the DrugBank database. We compared the coefficient of determination, root mean square error and found out that our new parameters outperform the previously published parameters. Coverage of the old parameters on this small evaluation dataset is like coverage on whole databases.

The quality of EEM parameters is also affected by a used QM charge scheme. EEM parameters derived from MPA, NPA, and AIM charges showed high quality. EEM parameters based on Hirshfeld charges were acceptable, and MK and CHELPG charges cannot be used with EEM. On the other hand, none of the QM theory level and basis set combinations showed any problem in the quality of EEM parameters. This also confirmed that we used an appropriate selection of reference QM charges.

We also evaluated already existing tools for calculating partial atomic charges, and all the tools showed some issues. For example, OpenBabel tool is using $\mathrm{Bult2002}_m pa parameterset$ 1001[?], $but developersextendedthisset aboutmissingatomtypeswithparameters fordiffer entatomtyp$

5.2.1 My contribution

I participated in the study's design, and I cooperated in the preparation of the input data (molecules and published EEM parameters) and in QM charge calculation. I performed the analyses and the interpretation of the data.

5.3 NEEMP: software for validation, accurate calculation, and fast parameterization of EEM charges

This article describes NEEMP – a software tool with three main functionalities – parametrization of EEM charges from reference QM charges, validation of EEM parameter sets (including quality and coverage calculation), and EEM charge calculation.

NEEMP provides two different parametrization approaches:

- linear regression (LE),
- differential evolution with the local minimization (DE-MIN) approach.

A combination of a global optimization method with a local optimization method improves EEM parameterization. This combined approach provides a more robust methodology than LR. Therefore it is applicable even for heterogeneous training sets. Specifically, we combined differential evolution (DE) with the local minimization method NEWUOA [47]. Quality criteria for evaluation of each iteration of the parametrization process (both LR and DE-MIN) can be set up to coefficient of determination (R^2) or the root mean square error (RMSE).

The validation mode of NEEMP calculates quality metrics, coverage, and a graphical representation of EEM charge correlation with reference QM charges.

The calculation mode of NEEMP provides a calculation of EEM charges using an input parameter set.

The article also presents two case studies to show the functionality and performance of NEEMP – a parametrization and a validation case study.

The parametrization case study targets a comparison of the parameterization method (LR vs. DE-MIN) and metrics for model validation (R^2 vs. RMSE). The case study proved that LR (with both metrics) is suitable for smaller and homogeneous datasets. DE-MIN (with RMSD metric) is a more robust approach that can also handle the parametrization of larger and more heterogeneous datasets. The validation case study provided similar findings to the previous article –low coverage of the older parameter sets. Also, a quality validation agrees with the previous article for smaller datasets with molecules comprised of C, H, N, and O. On the other hand, the case study uncovered an interesting problem: in larger and more heterogeneous datasets - the parameters set from our previous article proved accuracy problems with some atom types. Using NEEMP, we computed parameter sets, which are also applicable for such problematic datasets.

5.3.1 My contribution

I prototyped the DE-Min approach and designed some new NEEMP features, such as the usage of RMSE instead of R2. I also implemented a preparation of validion reports.

6

Follow-up work and future plans

Part IV

Conclusion

7

Conclusion

Appendix

Bibliography

[1] Berman, H. M., Kleywegt, G. J., Nakamura, H., and Markley, J. L. (2014) The protein data bank archive as an open data resource. *Journal of computer-aided molecular design*, **28**, 1009–1014.

[2] Bolton, E. E., Wang, Y., Thiessen, P. A., and Bryant, S. H. (2008) Pubchem: integrated platform of small molecules and biological activities. *Annual reports in computational chemistry*, **4**, 217–241.

[3] Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., and Coleman, R. G. (2012) Zinc: A free tool to discover chemistry for biology. *Journal of chemical information and modeling*, **52**, 1757–1768, pMID: 22587354.

[4] Gaulton, A., et al. (2012) Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, **40**, D1100–D1107.

[5] Paulsen, C. E., Armache, J.-P., Gao, Y., Cheng, Y., and Julius, D. (2015) Structure of the TRPA1 ion channel suggests regulatory mechanisms. *Nature*, **520**, 511–517.

[6] Cao, E., Liao, M., Cheng, Y., and Julius, D. (2013) Trpv1 structures in distinct conformations reveal activation mechanisms. *Nature*, **504**, 113–118.

[7] Prota, A. E., Bargsten, K., Zurwerra, D., Field, J. J., Díaz, J. F., Altmann, K.-H., and Steinmetz, M. O. (2013) Molecular mechanism of action of microtubule-stabilizing anticancer agents. *Science*, **339**, 587–590.

[8] Lu, J., et al. (2005) Microrna expression profiles classify human cancers. *Nature*, **435**, 834–838.

[9] Puente, X. S., et al. (2011) Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature*, **475**, 101–105.

[10] Nayal, M. and Honig, B. (2006) On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins*, **63**, 892–906.

[11] Xie, L., Xie, L., and Bourne, P. (2009) A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics*, **25**, i305–i312.

[12] Comer, J. and Tam, K. (2001) *Pharmacokinetic Optimization in Drug Research: Biological, Physicochemical, and Computational Strategies*. Verlag Helvetica Chimica Acta, Postfach, CH-8042 Zürich, Switzerland.

[13] Klebe, G. (2000) Recent developments in structure-based drug design. *Journal of molecular medicine*, **78**, 269–281.

[14] Kim, J. H., Gramatica, P., Kim, M. G., Kim, D., and Tratnyek, P. G. (2007) Qsar modelling of water quality indices of alkylphenol pollutants. *SAR and QSAR in environmental research*, **18**, 729–743.

[15] Gasteiger, J. and Engel, T. (2006) *Chemoinformatics: a textbook*. John Wiley & Sons.

[16] Atkins, P. and De Paula, J. (2011) *Physical chemistry for the life sciences*. Oxford University Press.

[17] Chaves, J., Barroso, J. M., Bultinck, P., and Carbo-Dorca, R. (2006) Toward an alternative hardness kernel matrix structure in the electronegativity equalization method (eem). *Journal of chemical information and modeling*, **46**, 1657–1665.

[18] Gross, K. C., Seybold, P. G., and Hadad, C. M. (2002) Comparison of Different Atomic Charge Schemes for Predicting pKa Variations in Substituted Anilines and Phenols. *International journal of quantum chemistry*, **90**, 445–458.

[19] Moller, H., Martinez-Yamout, M., Dyson, H., and Wright, P. (2005) Solution structure of the N-terminal zinc fingers of the Xenopus laevis double-stranded RNA-binding protein ZFa. *Journal of molecular biology*, **351**, 718–730.

[20] Zhang, J., Kleinöder, T., and Gasteiger, J. (2006) Prediction of pKa Values for Aliphatic Carboxylic Acids and Alcohols With Empirical Atomic Charge Descriptors. *Journal of chemical information and modeling*, **46**, 2256–2266.

[21] Ghafourian, T. and Dearden, J. (2000) The Use of Atomic Charges and Orbital Energies as Hydrogen-bonding-donor Parameters for QSAR Studies: Comparison of MNDO, AM1 and PM3 Methods. *Journal of pharmacy and pharmacology*, **52**, 603–610.

[22] Galvez, J., Garcia, R., Salabert, M. T., and Soler, R. (1994) Charge Indexes. New Topological Descriptors. *Journal of chemical information and modeling*, **34**, 520–525.

[23] Stalke, D. (2011) Meaningful structural descriptors from charge density. *Chemistry*, **17**, 9264–78.

[24] Lyne, P. D. (2002) Structure-based virtual screening: an overview. *Drug discovery today*, **7**, 1047–1055.

[25] Bissantz, C., Folkers, G., and Rognan, D. (2000) Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *Journal of medicinal chemistry*, **43**, 4759–4767.

[26] Mulliken, R. S. (1955) Electronic Population Analysis on LCAO-MO Molecular Wave Functions. II. Overlap Populations, Bond Orders, and Covalent Bond Energies. *Journal of chemical physics*, **23**, 1841.

[27] Mulliken, R. S. (1955) Electronic Population Analysis on LCAO-MO Molecular Wave Functions. I. *Journal of chemical physics*, **23**, 1833.

[28] Reed, A. E. and Weinhold, F. (1983) Natural bond orbital analysis of near-Hartree-Fock water dimer. *Journal of chemical physics*, **78**, 4066–4073.

[29] Reed, A. E., Weinstock, R. B., and Weinhold, F. (1985) Natural population analysis. *Journal of chemical physics*, **83**, 735.

[30] Löwdin, P.-O. (1950) On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals. *Journal of chemical physics*, **18**, 365.

[31] Hirshfeld, F. L. (1977) Bonded-atom fragments for describing molecular charge densities. *Theoretica chimica acta*, **44**, 129–138.

[32] Bader, R. F. W. (1985) Atoms in molecules. *Accounts of chemical research*, **18**, 9–15.

[33] Bader, R. F. W. (1991) A quantum theory of molecular structure and its applications. *Chemical reviews*, **91**, 893–928.

[34] Breneman, C. M. and Wiberg, K. B. (1990) Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *J. Comput. Chem.*, **11**, 361–373.

[35] Besler, B. H., Merz, K. M., and Kollman, P. A. (1990) Atomic charges derived from semiempirical methods. *Journal of computational chemistry*, **11**, 431–439.

[36] Marenich, A. V., Cramer, C. J., and Truhlar, D. G. (2009) 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. *Journal of physical chemistry B*, **113**, 6378–96.

[37] Cho, K.-H., Kang, Y. K., No, K. T., and Scheraga, H. A. (2001) A Fast Method for Calculating Geometry-Dependent Net Atomic Charges for Polypeptides. *Journal of physical chemistry B*, **105**, 3624–3634.

[38] Oliferenko, A. A., Pisarev, S. A., Palyulin, V. A., and Zefirov, N. S. (2006) Atomic Charges via Electronegativity Equalization: Generalizations and Perspectives. *Advances in quantum chemistry*, **51**, 139–156.

[39] Rappe, A. K. and Goddard, W. A. (1991) Charge equilibration for molecular dynamics simulations. *Journal of physical chemistry*, **95**, 3358–3363.

[40] Nistor, R. A., Polihronov, J. G., Müser, M. H., and Mosey, N. J. (2006) A generalization of the charge equilibration method for nonmetallic materials. *Journal of chemical physics*, **125**, 094108.

[41] Mortier, W. J., Ghosh, S. K., and Shankar, S. (1986) Electronegativity Equalization Method for the Calculation of Atomic Charges in Molecules. *Journal of the American Chemical Society*, **108**, 4315–4320.

[42] Ouyang, Y., Ye, F., and Liang, Y. (2009) A modified electronegativity equalization method for fast and accurate calculation of atomic charges in large biological molecules. *Physical chemistry chemical physics*, **11**, 6082–9.

[43] Ishihama, Y., Nakamura, M., Miwa, T., Kajima, T., and Asakawa, N. (2002) A rapid method for pk_a determination of drugs using pressure-assisted capillary electrophoresis with photodiode array detection in drug discovery. *Journal of pharmaceutical sciences*, **91**, 933–942.

[44] Babić, S., Horvat, A. J., Pavlović, D. M., and Kaštelan-Macan, M. (2007) Determination of pk_a values of active pharmaceutical ingredients. *TrAC*, **26**, 1043–1061.

[45] Manallack, D. (2007) The pk$_a$ distribution of drugs: Application to drug discovery. *Perspectives in medicinal chemistry*, **1**, 25–38.

[46] Wan, H. and Ulander, J. (2006) High-throughput pk_a screening and prediction amenable for adme profiling. *Expert opinion on drug metabolism & toxicology*, **2**, 139–155.

[47] Cruciani, G., Milletti, F., Storchi, L., Sforna, G., and Goracci, L. (2009) *In silico* pK_a prediction and adme profiling. *Chemistry & biodiversity*, **6**, 1812–1821.

[48] Lee, A. C. and Crippen, G. M. (2009) Predicting pk_a. *Journal of chemical information and modeling*, **49**, 2013–2033.

[49] Rupp, M., Körner, R., and Tetko, I. V. (2010) Predicting the pK$_a$ of small molecules. *Combinatorial chemistry and high throughput screening*, **14**, 307–327.

[50] Fraczkiewicz, R. (2006) *In Silico Prediction of Ionization*, vol. 5. Elsevier.

[51] Ho, J. and Coote, M. (2010) A universal approach for continuum solvent pk_a calculations: Are we there yet? *Theoretica chimica acta*, **125**, 3–21.

[52] Clark, J. and Perrin, D. D. (1964) Prediction of the strengths of organic bases. *Quarterly reviews of the Chemical Society*, **18**, 295–320.

[53] Perrin, D. D., Dempsey, B., and Serjeant, E. P. (1981) *pK_a prediction for organic acids and bases*. Chapman and Hall: New York.

[54] Blower, P. E. and Cross, K. P. (2006) Decision tree methods in pharmaceutical research. *Current topics in medicinal chemistry*, **6**, 31–39.

[55] Liptak, M. D., Gross, K. C., Seybold, P. G., Feldgus, S., and Shields, G. (2002) Absolute pk_a determinations for substituted phenols. *Journal of the American Chemical Society*, **124**, 6421–6427.

[56] Toth, A. M., Liptak, M. D., Phillips, D. L., and Shields, G. C. (2001) Accurate relative pk_a calculations for carboxylic acids using complete basis set and gaussian-n models combined with continuum solvation methods. *Journal of chemical physics*, **114**, 4595–4606.

[57] Citra, M. J. (1999) Estimating the pk_a of phenols, carboxylic acids and alcohols from semi-empirical quantum chemical methods. *Chemosphere*, **1**, 191–206.

[58] Kreye, W. C. and Seybold, P. G. (2009) Correlations between quantum chemical indices and the pk_as of a diverse set of organic phenols. *International journal of quantum chemistry*, **109**, 3679–3684.

Appendix: Main papers

Appendix: All My Publications

Curriculum Vitae