Multiple-Choice Knapsack for Assigning Partial Atomic Charges in Drug-Like Molecules

Martin S. Engler
Life Sciences and Health Group, Centrum Wiskunde & Informatica,
Amsterdam, The Netherlands
martin.engler@cwi.nl

Bertrand Caron
School of Chemistry & Molecular Biosciences,
The University of Queensland, St Lucia, Australia
https://orcid.org/0000-0003-2305-1452

Lourens Veen
Netherlands eScience Center,
Amsterdam, The Netherlands

Daan P. Geerke
AIMMS Division of Molecular and Computational Toxicology,
Vrije Universiteit Amsterdam, The Netherlands
https://orcid.org/0000-0002-5262-6166

Alan E. Mark
School of Chemistry & Molecular Biosciences,
The University of Queensland, St Lucia, Australia
https://orcid.org/0000-0001-5880-4798

Gunnar W. Klau
Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, Germany
gunnar.klau@hhu.de
https://orcid.org/0000-0002-6340-0090

Abstract

A key factor in computational drug design is the consistency and reliability with which intermolecular interactions between a wide variety of molecules can be described. Here we present a procedure to efficiently, reliably and automatically assign partial atomic charges to atoms based on known distributions. We formally introduce the molecular charge assignment problem, where the task is to select a charge from a set of candidate charges for every atom of a given query molecule. Charges are accompanied by a score that depends on their observed frequency in similar neighbourhoods (chemical environments) in a database of previously parameterised molecules. The aim is to assign the charges such that the total charge equals a known target charge within a margin of error while maximizing the sum of the charge scores. We show that the problem is a variant of the well-studied multiple-choice knapsack problem and thus weakly $\mathcal{NP}$-complete. We propose solutions based on Integer Linear Programming and a pseudo-polynomial time Dynamic Programming algorithm. We show that the results obtained for novel molecules not included in the database are comparable to the ones obtained performing explicit charge calculations while decreasing the time to determine partial charges for a molecule by several orders of magnitude, that is, from hours or even days to below a second.

Our software is openly available at `https://github.com/enitram/charge_assign`.

2012 ACM Subject Classification Applied computing → Chemistry

18th International Workshop on Algorithms in Bioinformatics (WABI 2018).
Editors: Laxmi Parida and Esko Ukkonen; Article No. 16; pp. 16:1–16:13

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Keywords and phrases Multiple-choice knapsack, integer linear programming, pseudo-polynomial dynamic programming, partial charge assignment, molecular dynamics simulations

Digital Object Identifier 10.4230/LIPIcs.WABI.2018.16

Funding This research was partially supported by the Netherlands eScience Center (NLeSC), grant number 027.015.G06.

Acknowledgements We thank all members of the NLeSC-ASDI project *Enhancing Protein-Drug Binding Prediction* for valuable discussions. We thank Ulrich Pferschy from the University of Graz for providing insight and expertise on the multiple-choice knapsack problem.

1 Introduction

Molecule-based computational modelling and simulation studies play a central role in modern drug design and development. In particular, molecular dynamics (MD) simulations and free energy calculations are increasingly being used to screen potential ligand molecules in terms of their interactions with proposed target molecules (e.g. cell surface receptors or enzymes involved in metabolism) [11, 18]. They are also used to model structural changes in the target molecule associated with the binding of a given drug in order to understand the mechanism of action. The accuracy and utility of such modelling studies depends directly on the fidelity with which intermolecular interactions can be represented [1, 15]. While ideally one might wish to represent such interactions on the level of quantum mechanics, the size and complexity of protein/ligand complexes necessitates the use of classical dynamics in conjunction with empirical potentials. These so-called force fields are parameterised to reproduce the interactions between atoms in a system of interest (e.g. protein, membrane, drug) and involve bonds, angles, dihedrals, van der Waals and coulombic interactions.

Of particular importance is the assignment of partial atomic charges to describe the latter interactions. Partial atomic or point charges are used to represent the electrostatic potential around a molecule and the coulombic interactions between these point charges dominate the calculation of inter-molecular interactions. The difficulty is that the effective partial charge on an atom needed to represent the electrostatic potential surrounding a molecule is heavily dependent on the local environment in which an atom is found. For small molecules (< 40 atoms) partial atomic charges can be generally inferred *de novo* from quantum-mechanical computations [19]. However, when using e.g. commonly applied Density Functional Theory (DFT) such calculations scale cubic in the number of valence electrons [3], increasing the computational costs significantly. In addition, as molecules become larger the accuracy with which charges can be assigned decreases.

The standard approach to address this problem is to manually assign charges to atoms based on their similarity to atoms (or groups) in a set of reference molecules containing equivalent chemical moieties. The challenge in making such assignments is twofold: 1) the charges assigned to equivalent chemical groups in alternative reference molecules may vary making the choice of a reference molecule difficult and 2) the charges assigned to neighbouring atoms must be consistent. In particular, the total charge on the molecule must be integer. In the recent years, a number of machine learning approaches emerged that infer charges based on a set of reference molecules [4, 13, 16]. However, these approaches often struggle to deal with the ambiguity of similar groups that have different charges in different molecules and the requirement that the overall charge must be integer.

In this paper, we consider the problem of – given a large set of reference molecules with known charge distributions – how to efficiently, automatically and optimally assign partial atomic charges which are consistent with both the neighbouring atoms and the total charge. As a reference we have used molecules parameterised using the Automated Topology Builder (ATB) and repository [15]. The ATB contains a large number of molecules (< 50 atoms) for which partial charges have been assigned *de novo*. In previous work, we have contributed to improving the reliability of this repository by ensuring the consistency and utility of the partial charges assigned to atoms by identifying atoms that could be used to form *charge groups*, which can be collectively assigned integer formal charges $(\ldots, -1, 0, 1, \ldots)$ [5]. We have also developed methods to match molecular substructures, taking into account that the partial charge of an atom is heavily dependent on its neighbours and the nature of its local chemical environment [7]. This made it possible to study the distribution of charges within local molecular environments for all molecules in the ATB ($\approx$ 200,000 molecules; 7,800,000 atoms and 5,600,000 bonds) and to find, given a query molecule, all possible matching fragments (sub-graphs).

Here we build on this previous work and our ability to match sub-fragments of a query molecule against the available database, to consider how the information contained in already parameterised molecules can be used best to infer the charges within a novel molecule. The most direct approach would be to simply use the mean partial charge on individual atoms identified as equivalent using a given similarity criterion. However, quantum mechanics dictates that the total charge on a molecule must be integer. Simply attributing to each atom the value of the mean partial charge from the known distribution fails as it results in the accumulation of errors and a total charge deviating from the required value.

Instead, we have considered solutions to the *molecular charge assignment problem*, which allow charges that deviate from the mean to be selected while their sum is constrained to lie close to a target total charge. Among the possible set of solutions we prefer those that maximize a score that depends on the observed frequencies of the chosen charges. We show that the problem is similar to a *multiple-choice knapsack problem* (MCKP) [6, 14]. We introduce ϵ-MCKP, a variant of the standard MCKP with an error margin ϵ. We provide an Integer Linear Programming (ILP) formulation of ϵ-MCKP and adapt the MCKP pseudo-polynomial Dynamic Programming (DP) algorithm to ϵ-MCKP. Finally, we demonstrate the utility of the ϵ-MCKP approach for solving the charge assignment problem by comparing the charges proposed based on this approach to those obtained directly using the ATB. We find that the charges computed with our novel approach are comparable to the ones obtained using explicit charge calculations while decreasing the time to determine partial charges for a molecule by several orders of magnitude, that is, from hours or even days to below a second.

Our code is publicly available at `https://github.com/enitram/charge_assign` under the Apache 2.0 open source license.

2 Assigning charges

We consider molecules as graphs. Let $G = (V, E, t)$ be a *molecular graph*, where vertices V correspond to atoms, edges E correspond to bonds and $t : V \to \Sigma$ colors vertices with atom types. A straightforward alphabet of atom types Σ would be the chemical elements. In this work, we used ATB-assigned GROMOS atom types which provide a more detailed classification of some chemical elements (N, C, O, S) depending on their hybridization (number of bonded nodes), and therefore provides a more detailed description of the local environment.

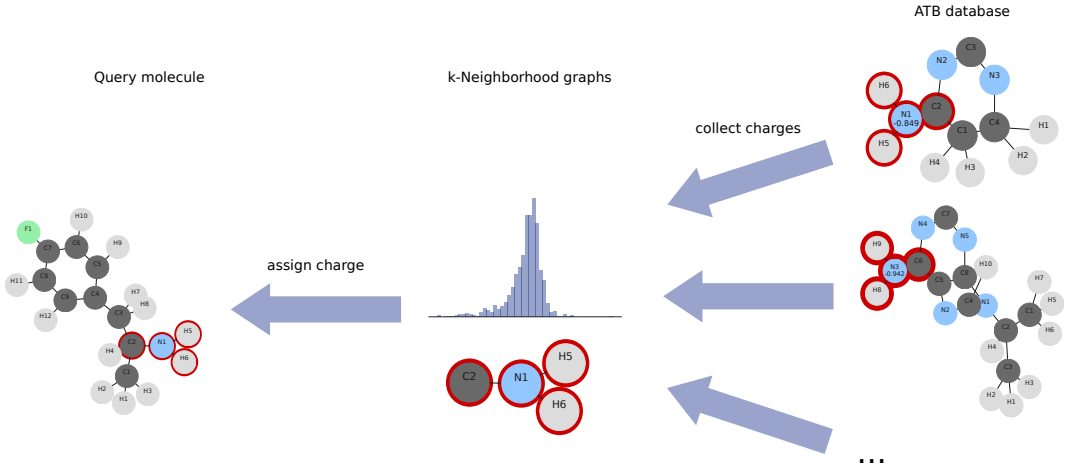

Figure 1 Given a query molecule, our method assigns atomic partial charges based on matching isomorphic subgraphs (red) with a known partial charge distribution collected from the ATB database of parameterised molecules.

The partial charge of an atom is heavily dependent on its bonded neighbours and the nature of its local environment. Formally, we define the k-neighbourhood as:

▶ **Definition 1** (k-neighbourhood)**.** Let $N(v) = \{u \mid (u, v) \in E\}$ be the neighbourhood of an atom v. We define the *k-neighbourhood* recursively as $N_k(v) = N_{k-1}(v) \cup \bigcup_{u \in N_{k-1}(v)} N(u)$, with $N_0(v) = v$.

Informally, the k-neighbourhood of an atom v is the set of all atoms for which a path of length $\leq k$ to v exists. Let $G[N_k(v)]$ be the subgraph induced by the k-neighbourhood of v.

To collect all possible partial charge values, we consider all k-neighbourhoods in the set of previously parameterised molecules. For this we iterate over all atoms v of all molecular graphs in the ATB and construct a list of subgraphs $G[N_k(v)]$ with associated partial charges of the corresponding atom v. We construct a database with an entry for each isomorphism class in the subgraph list. For each isomorphism class we collect the partial charges of its subgraphs and condense the values to a histogram. Since the point charges assigned by the ATB are rounded to three digits after the decimal point, we round the partial charge values accordingly.

Given a query molecule with a known target total charge, the challenge is to assign the most representative partial charge to each atom while staying close to the target total charge (Fig. 1). For that purpose, we iterate over all atoms of the query molecular graph and generate the subgraphs $G[N_k(v)]$. We match each subgraph to its isomorphism class in our database of k-neighbourhood subgraphs. If there is no match, we iteratively retry with $G[N_{k-1}(v)]$ until $k = 0$. Now each atom in our query molecule has a histogram of possible partial charges. The task is now to assign the charges such that we maximize the frequencies of the assigned charges while the sum of assigned partial charges equals the target charge with some error margin.

3 Problem Formulation and Complexity

We map each atom i to a set of items j with weights $w_{i,j}$ corresponding to partial charges and profits $p_{i,j}$ corresponding to their frequency-based scores. The target total charge corresponds to capacity c. Note that the charge assignment problem is now similar to a multiple-choice knapsack problem (MCKP). The decision version of MCKP is defined as:

▶ **Problem 1** (MCKP). *Given a decision variable $K \geq 0$, capacity $c \geq 0$, m sets $N_1, \ldots, N_m$ of items $j \in N_i$ with profit $p_{i,j} \geq 0$ and weight $w_{i,j} \geq 0$, select exactly one item from each set, such that the sum of weights of the selected items does not exceed c and the sum of profits of the selected items is equal or larger than K.*

MCKP is known to be weakly $\mathcal{NP}$-complete [6, 9, 14]. However, although the problem of assigning charges is similar to MCKP, there are two differences. First, weights and capacity can be negative numbers. Second, the sum of weights of selected items must hit the capacity with some error margin, resulting in an upper and lower capacity limit. We define a variant of MCKP, which is equivalent to the charge assignment problem as:

▶ **Problem 2** (ϵ-MCKP). *Given a decision variable $K \geq 0$, capacity $-\infty \leq c \leq \infty$, error $\epsilon \geq 0$, m sets $N_1, \ldots, N_m$ of items $j \in N_i$ with profit $p_{i,j} \geq 0$ and weight $-\infty \leq w_{i,j} \leq \infty$, select exactly one item from each set, such that the sum of weights of the selected items is in the range $[c - \epsilon, c + \epsilon]$ and the sum of profits of the selected items is equal or larger than K.*

▶ **Theorem 2.** *ϵ-MCKP is weakly $\mathcal{NP}$-complete.*

Proof. Showing that ϵ-MCKP is in $\mathcal{NP}$ is straightforward. Given an instance of ϵ-MCKP and a candidate solution $\hat{S}$, we can easily check in polynomial time whether $c - \epsilon \leq \sum_{w_{i,j} \in \hat{S}} w_{i,j} \leq c + \epsilon$ and $\sum_{p_{i,j} \in \hat{S}} p_{i,j} \geq K$ as well as if $\hat{S}$ contains exactly one item from each set $N_1, \ldots, N_m$. We show that ϵ-MCKP is weakly $\mathcal{NP}$-hard as follows: We reduce MCKP $\leq_{\mathrm{P}}$ ϵ-MCKP. Given an instance of the standard MCKP with capacity c, we transform it to an ϵ-MCKP instance with capacity $c' = \frac{1}{2}c$ and $\epsilon = \frac{1}{2}c$. Then, $c' - \epsilon = 0$ and $c' + \epsilon = c$, making both instances equivalent. ◀

Both problems obviously can be transformed into optimization problems by omitting the decision variable K and maximizing the sum of profits. The definition of ϵ-MCKP allows us to solve the charge assignment problem.

4 Solving ϵ-MCKP

In this section we present two algorithmic strategies to solve ϵ-MCKP: the first is based on an integer linear programming (ILP) formulation, which can be solved by general ILP solvers, while the second is a purely combinatorial dynamic programming (DP) algorithm.

Formulating ϵ-MCKP as an ILP is straightforward. Let $x_{i,j}$ be a binary variable with value 1 if and only if item j in set N_i is selected. We formulate the problem as:

$$\max \quad \sum_{i=1}^{m} \sum_{j \in N_i} x_{i,j} p_{i,j} \tag{1a}$$

$$\text{subject to} \quad \sum_{i=1}^{m} \sum_{j \in N_i} x_{i,j} w_{i,j} \geq c - \epsilon \tag{1b}$$

$$\sum_{i=1}^{m} \sum_{j \in N_i} x_{i,j} w_{i,j} \leq c + \epsilon \tag{1c}$$

$$\sum_{j \in N_i} x_{i,j} = 1 \qquad \text{for } 1 \leq i \leq m \tag{1d}$$

$$x_{i,j} \in \{0, 1\} \qquad \text{for } 1 \leq i \leq m, j \in N_i \tag{1e}$$

The second algorithm is an adaption of the pseudo-polynomial DP of the standard MCKP to ϵ-MCKP. The standard MCKP assumes numbers to be non-negative integers. If a given

ϵ-MCKP instance does not comply with the non-negativity and integrality constraints, we transform the instance as follows:

First, we convert floating point weights $w_{i,j}$, capacity c and error ϵ to integers by multiplying with an appropriate factor. Since point charges in this work are rounded to three digits after the decimal point, a factor of 10^3 is sufficient. Second, we transform the weights $w_{i,j}$ and capacity c to non-negative numbers. For every set N_j with $j = 1, \ldots, m$, we determine the minimum weight $w_i^* = \min_{j \in N_i} w_{i,j}$. We define the new weights as $\tilde{w}_{i,j} = w_{i,j} - w_i^*$. Then, the weights are guaranteed to be non-negative. As we have to select one item per set, we can define the new capacity as $\tilde{c} = c - \sum_{i=1}^{m} w_i^*$.

Therefore, we assume in the following (without loss of generality) that weights $w_{i,j}$, capacity c and error ϵ are non-negative integers. Let P be a two-dimensional DP-table of size $m \times (c + \epsilon)$. $P[k, d]$ holds the maximum profit that we can achieve with sets 0 to k and a sum of weights of exactly d:

$$P[k,d] = \max \left\{ \sum_{i=0}^{k} \sum_{j \in N_i} x_{i,j} p_{i,j} : \sum_{i=0}^{k} \sum_{j \in N_i} x_{i,j} w_{i,j} = d, \sum_{j \in N_k} x_{i,j} = 1 \text{ for all } 0 \leq i \leq k \right\} \quad (2)$$

We compute P recursively. Let $P[k, d]$ be defined as:

$$P[k,d] = \max \begin{cases} P[k-1, d - w_{k,j}] + p_{k,j} & \text{for } j \in N_i \text{ and } d - w_{k,j} \geq 0 \\ -\infty \end{cases} \quad (3)$$

$P[k, d]$ is calculated by considering all items of the current set N_k and computing the maximum profit that can be achieved when adding those profits to possible previous solutions with $k - 1$ sets and sum of weights $d - w_{k,j}$. The profit is $-\infty$ if there is no possible solution for $P[k, d]$. Contrary to the standard MCKP DP we initialize P as:

$$P[0,d] = \begin{cases} 0 & \text{if } d = 0 \\ -\infty & \text{otherwise} \end{cases} \quad (4)$$

This ensures that only solutions in which the sum of selected weights equals exactly d are possible. We find the maximum profit p^* by:

$$p^* = \max \{ P[m,d] : \max\{c - \epsilon, 0\} \leq d \leq c + \epsilon \} \quad (5)$$

The DP can be easily implemented using one dimension, as the recursion only looks back one step in the dimension k (the number of sets we currently consider). The space requirement of the DP algorithm is $O(c + \epsilon)$. The running time complexity is $O(n(c + \epsilon))$, with n being the total number of items.

5 Score

We modeled the charge assignment problem as ϵ-MCKP, where atoms i are sets of items j with weight $w_{i,j}$ corresponding to partial charges and profits $p_{i,j}$ corresponding to their scores. In this section we propose a frequency-based score for the ϵ-MCKP profit maximization.

Figure 2 shows the distribution of charges over all 3-neighbourhood graphs in a snapshot of the ATB of roughly 160,000 molecules centered at the sample mean of each 3-neighbourhood graph. At a first glance, it may seem to be normally distributed, but the Q-Q-plot on the right hand side of Figure 2 reveals that the distribution is heavy-tailed. Therefore, using measures that assume normally distributed data such as the z-score is not advisable. We also

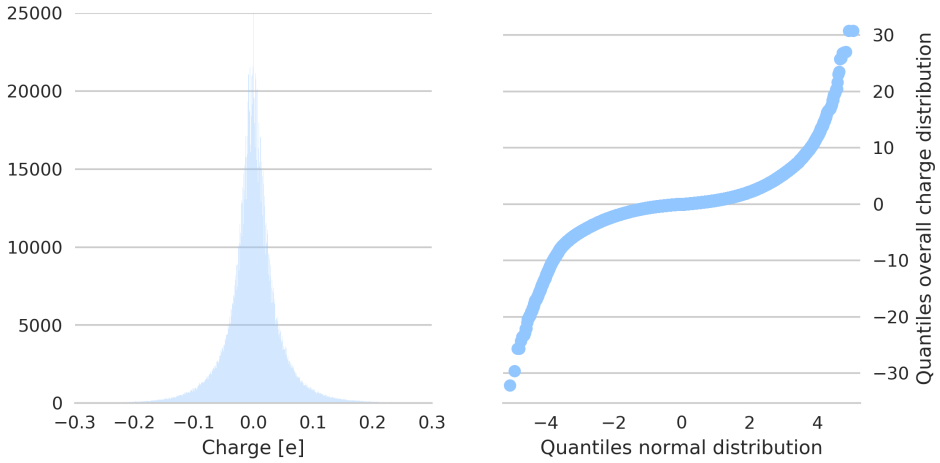

Figure 2 Distribution of charges over all 3-neighbourhood graphs centered at the sample mean of each 3-neighbourhood graph (left) and Q-Q-plot with the quantiles of the charge distribution over all 3-neighbourhood graphs on the y-axis and the quantiles of a fitted normal distribution on the x-axis.

refrain from simply using the logarithm of the frequencies as our score, since the deviation of the sample mean of the observed charges should also be taken into account. Additionally, the logarithm will result in negative profits.

We propose a simple score using squared distances. Let $f_{i,j}$ be the observed frequency of partial charge $w_{i,j}$ and $\hat{\mu}_i$ the sample mean of all observed charges of atom i. We define the score $p_{i,j}$ as

$$p_{i,j} = \frac{f_{i,j}}{1 + (w_{i,j} - \hat{\mu}_i)^2}. \tag{6}$$

The score reflects both the observed frequency of a partial charge and its distance to the observed mean of all partial charges of an atom. If the charge equals the mean $w_{i,j} = \hat{\mu}_i$, then the score equals its observed frequency $p_{i,j} = f_{i,j}$. The larger the distance of a charge to the mean is, the smaller the score will be. This serves as a tie-breaker, such that if two charges have the same observed frequency and are within the capacity limits, ϵ-MCKP will prefer the charge closer to the observed mean.

6 Results and Discussion

To evaluate our method, we conducted a leave-one-out-analysis using a snapshot of the ATB database containing roughly 160,000 molecules. We focus on this set of previously computed molecules, since the computational effort of large-scale quantum-mechanical calculations is significant. We created a database of k-neighbourhood subgraphs associated with partial charge histograms with variable bin widths and a fixed $k = 3$. Bin widths were determined according to the Friedman-Diaconis rule [8]. Then, we temporarily removed all charge values associated with its 3-neighbourhood subgraphs for each molecule and computed the atomic partial charges using the new, smaller histogram database. We compared the assigned values to the original atomic partial charges in the ATB database.

All computations were performed on a compute cluster with 16 3.2 GHz Xeon CPUs and 512GB RAM. The ILP was solved using COIN-OR [17]. We recorded the running times of the ILP and DP algorithm, see Fig. 3. As expected, the running time of the DP scales

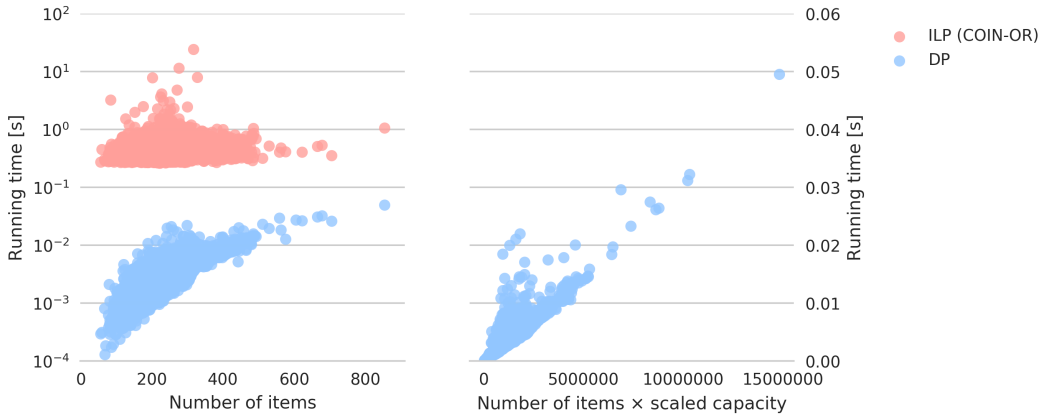

Figure 3 Running times of the ILP solved with COIN-OR and the DP dependent on the number of items n, showing that the DP is significantly faster than the ILP (left). The running time of the DP actually depends on the number of items times the scaled capacity $10^3 \cdot n \cdot (\tilde{c} + \epsilon)$ (right).

linearly with $10^3 \cdot n \cdot (\tilde{c} + \epsilon)$, where n is the number of items, 10^3 is the blowup factor and $\tilde{c}$ is the capacity c transformed to non-negativity with $\tilde{c} = c - \sum_{i=1}^{m} w_i^*$ and $w_i^* = \min_{j \in N_i} w_{i,j}$. The running times of the ILP show more variation and a marginal positive correlation to the number of items n, which equals the number of variables in the ILP. The DP was always significantly faster than the ILP in the leave-one-out-evaluation.

In the leave-one-out evaluation, we compared the naive approach of estimating the atomic partial charges by simply taking the mean and our method of solving an ϵ-MCKP instance, see Fig. 4. As expected, while the naive method on average is able to find charges with a slightly lower distance to the original partial charges, it often results in a total charge far away from the target total charge (with errors more than 1e in many cases). Our method on the other hand is able to assign charge values which are only slightly worse than the ones computed by the naive method while achieving a total charge close to the target total charge. As can be seen in Fig. 4 the deviation of the total charge from the target charge using the ϵ-MCKP approach is so small it is barely visible on the scale used.

As an example of the charges assigned by our method, Fig. 5 shows the two molecules with the atomic partial charges that are on average closest to and farthest from the original ATB charges. The computed charges for the molecule with the closest distance fit well to the original ATB charges. ϵ-MCKP assigns identical charges to atoms H1, H2 and H3. The 3-neighbourhood graphs of all three atoms have the same isomorphism class. This is an advantage of our ϵ-MCKP approach, since quantum-mechanical *de novo* charge assignment does not guarantee that similar charges are assigned to equivalent atoms (although in this case the ATB charges are also identical). For the molecule with the farthest distance there are some large distances of more than 1e. However, we observe that the large distances are caused by the original ATB charges being on the outer edges of the charge distributions, while ϵ-MCKP on the other hand picks charges close to the largest mode of the distribution, see bottom side of Fig. 5. Note that Fig. 5 shows the distributions used in the leave-one-out evaluation without the original ATB charges of the depicted molecule. Additionally, the charge distributions of atoms with large charge distances have been computed with a low number of observed charges, resulting in multimodal distributions with several large peaks. We expect this effect to disappear when more data is available in the constantly growing ATB repository.

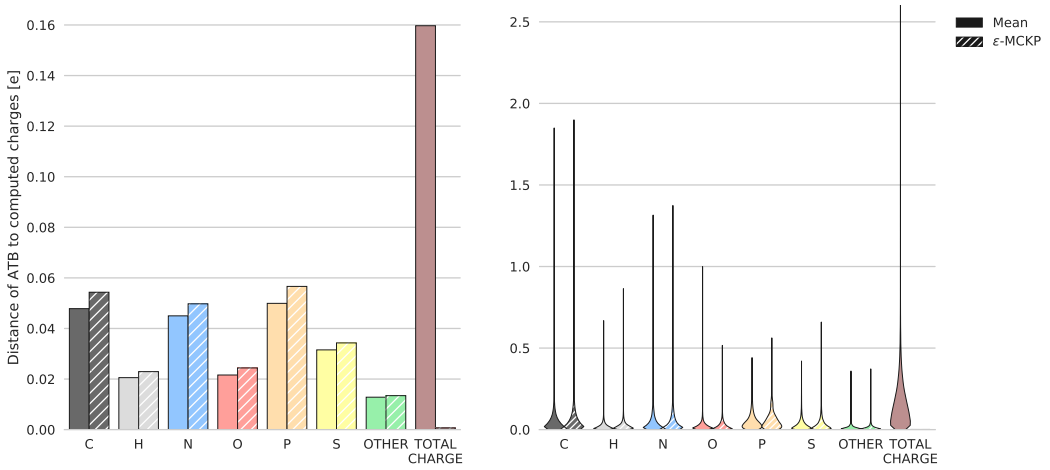

Figure 4 Results of the leave-one-out experiment with $k = 3$ showing mean distances in elementary charge units (left) and violin plots of all distances (right) of original charges found in the ATB to charges calculated by selecting the mean and solving ϵ-MCKP. The distances are categorized by chemical elements. For ϵ-MCKP, the computed total charges are virtually the same as the target total charges from the ATB, resulting in mean distances of almost zero.

Fig. 6 shows the chemical structure of a more complex example (ATB ID 25338). For this molecule, the *de novo* electrostatic-potential based charge assignment using quantum-mechanical computations required $\sim$ 140 days using on one core while solving our ϵ-MCKP approach was finished in $\sim$ 0.12 (ILP) and $\sim$ 0.06 (DP) seconds.

Most of the outer atoms – especially the hydrogens – showed a narrow unimodal distribution of ATB charges and ϵ-MCKP picked charges close to the original ATB charge. The more buried atoms showed a higher variability. For some atoms, we observed a similar behavior as in Fig. 5, that ϵ-MCKP selects a charge closer to the distribution mean than the original ATB charge was. However, for several atoms we observed the limit of our data-driven approach. If only a few charge values are available for a certain k-neighbourhood, then the distributions are multimodal with very similar or equal peak heights, reflecting the variability of the quantum-mechanically derived charges. Then, ϵ-MCKP may freely choose between co-optimal solutions. On the other hand, if only exactly one charge value is available, ϵ-MCKP has to choose this value. While the probability of this occurring will decrease with the addition of more data, in this case (with the current dataset) it would be advisable to use a smaller k. With choosing an appropriate k, the user may balance the specificity of large k-neighbourhoods against the robustness of small k-neighbourhoods.

In general, charges on the outer atoms of a molecule can be assigned quite well while charges of the inner atoms deviate more from the ATB charges. This may be explained by the higher variability of the inner atoms in the ATB dataset, an artifact of the *de novo* electrostatic-potential based charge assignment [2].

7 Conclusions

The ability to accurately calculate the electrostatic interactions between a ligand and its receptor is a key component of computer-aided drug development. In this paper, we have investigated the problem of automatically assigning partial charges. The charge assignment problem is similar to the multiple-choice knapsack problem. We introduced a variant tailored

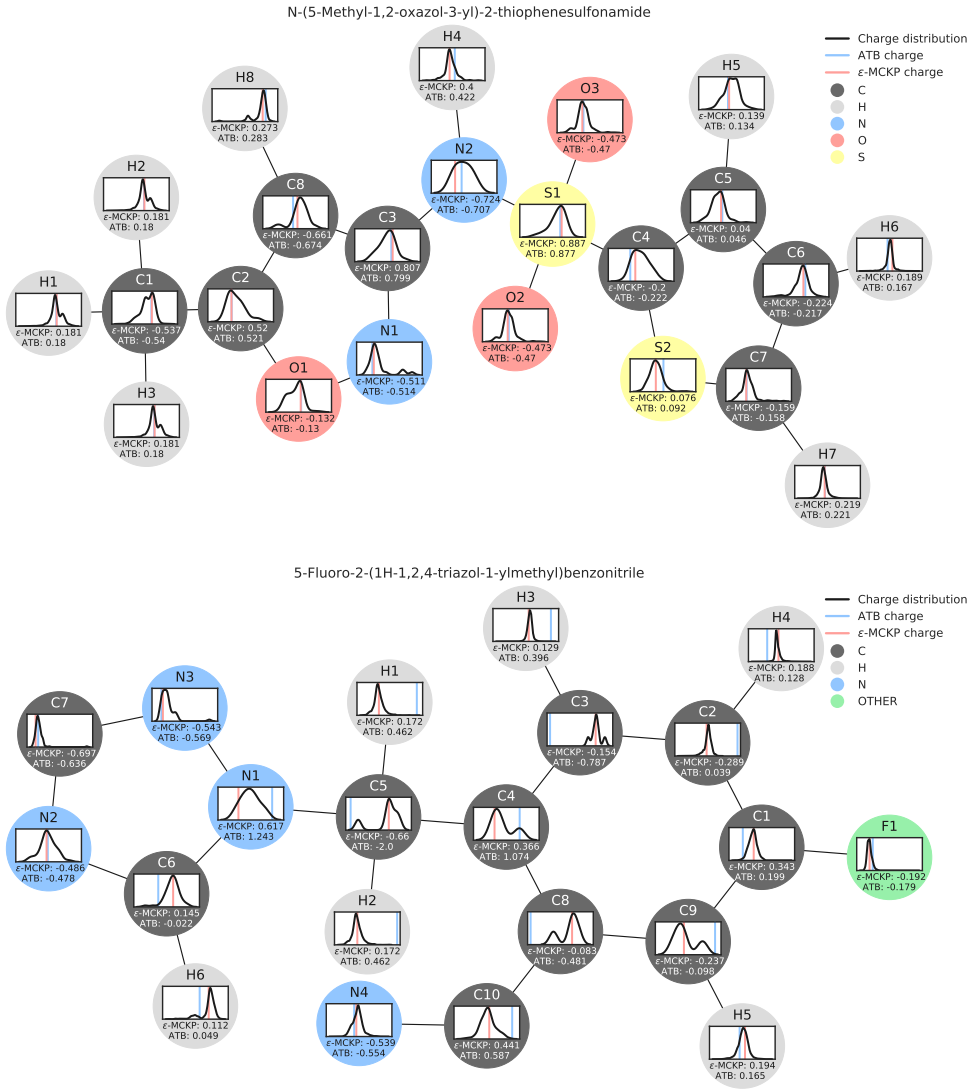

Figure 5 Best (top) and worst (bottom) molecules in the leave-one-out evaluation ranked by average distance of the original ATB charges to the charges computed by ϵ-MCKP. Atoms (nodes) are color-coded by their chemical element (red for oxygen, blue for nitrogen, black for carbon, yellow for sulfur and grey for hydrogen). Atoms are overlayed with the kernel-density estimate of the histograms of the leave-one-out charges of their respective 3-neighbourhoods. The original ATB charge and the computed charges are shown by blue and red vertical lines in the histograms.

to the charge assignment problem, the ϵ-multiple-choice knapsack problem (ϵ-MCKP). Like most knapsack problems, ϵ-MCKP is weakly $\mathcal{NP}$-complete. We presented two algorithmic solutions to ϵ-MCKP, an integer linear programming (ILP) formulation and a dynamic programming (DP) algorithm.

We conducted a leave-one-out evaluation on a snapshot of the ATB database. The computed atomic partial charges were close to the original ATB charges and the total charge virtually the same as the target total charge, suggesting that our method provides consistent parameters for MD simulations, docking studies and other related applications.

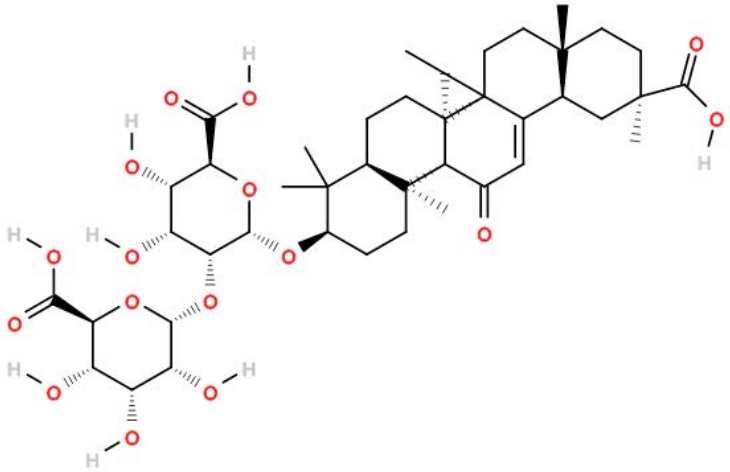

Figure 6 The chemical structure of ATB ID 25338 containing 120 atoms. Note that aliphatic hydrogens are not shown.

One additional advantage of our approach is that equivalent nodes in the graph will be assigned similar charges and the charge distribution will therefore mirror the symmetry of the molecular graph.

The DP algorithm performed faster than the ILP on a set of 160,000 molecules contained within the ATB. On average, both implementations required only a fraction of a second to assign charges to molecules containing 50-100 atoms, while quantum-mechanical computations required many days. This is important when screening large molecular databases. For instance, ChEMBL [10], a manually curated chemical database of bioactive molecules with drug-like properties, contains in excess of 1.6 million compounds. The majority of these have more than 50 atoms making quantum-mechanical computations difficult. Other computational drug design databases are larger again [20]. For example, ZINC, a free database of commercially-available compounds, contains more than 35 million compounds [12].

Our method builds on a repository of previously computed molecular parameters and assigns consistent partial atomic charges in a swift manner to facilitate MD simulations and related applications in drug design.

References

1 Robert Abel, Lingle Wang, David L. Mobley, and Richard A. Friesner. A critical review of validation, blind testing, and real-world use of alchemical protein-ligand binding free energy calculations. *Current Topics in Medicinal Chemistry*, 17(23):2577–2585, 2017. doi:10.2174/1568026617666170414142131.

2 Christopher I. Bayly, Piotr Cieplak, Wendy Cornell, and Peter A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges:

the RESP model. *The Journal of Physical Chemistry*, 97(40):10269–10280, 1993. doi:10.1021/j100142a004.

3 F. Matthias Bickelhaupt and Evert Jan Baerends. Kohn-Sham Density Functional Theory: Predicting and Understanding Chemistry. In Kenny B. Lipkowitz and Donald B. Boyd, editors, *Reviews in Computational Chemistry*, pages 1–86. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2007. doi:10.1002/9780470125922.ch1.

4 Patrick Bleiziffer, Kay Schaller, and Sereina Riniker. Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations. *Journal of Chemical Information and Modeling*, 58(3):579–590, 2018. doi:10.1021/acs.jcim.7b00663.

5 Stefan Canzar, Mohammed El-Kebir, René Pool, Khaled Elbassioni, Alpeshkumar K. Malde, Alan E. Mark, Daan P. Geerke, Leen Stougie, and Gunnar W. Klau. Charge Group Partitioning in Biomolecular Simulation. *Journal of Computational Biology*, 20(3):188–198, 2013. doi:10.1089/cmb.2012.0239.

6 Krzysztof Dudziński and Stanisław Walukiewicz. Exact methods for the knapsack problem and its generalizations. *European Journal of Operational Research*, 28(1):3–21, 1987. doi:10.1016/0377-2217(87)90165-2.

7 Martin S. Engler, Mohammed El-Kebir, Jelmer Mulder, Alan E. Mark, Daan P. Geerke, and Gunnar W. Klau. Enumerating common molecular substructures. *PeerJ Preprints*, 5:e3250v1, 2017. doi:10.7287/peerj.preprints.3250v1.

8 David Freedman and Persi Diaconis. On the histogram as a density estimator: L_2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57(4):453–476, 1981. doi:10.1007/BF01025868.

9 Michael R. Garey and David S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1990.

10 Anna Gaulton, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P. Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1):D1100–D1107, 2012. doi:10.1093/nar/gkr777.

11 Alexander Hillisch, Nikolaus Heinrich, and Hanno Wild. Computational chemistry in the pharmaceutical industry: From childhood to adolescence. *ChemMedChem*, 10(12):1958–1962, 2015. doi:10.1002/cmdc.201500346.

12 John J. Irwin and Brian K. Shoichet. ZINC - a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*, 45(1):177–182, 2005. doi:10.1021/ci049714+.

13 Maxim V. Ivanov, Marat R. Talipov, and Qadir K. Timerghazin. Genetic Algorithm Optimization of Point Charges in Force Field Development: Challenges and Insights. *The Journal of Physical Chemistry A*, 119(8):1422–1434, 2015. doi:10.1021/acs.jpca.5b00218.

14 Hans Kellerer, Ulrich Pferschy, and David Pisinger. *Knapsack problems*. Springer Berlin, Berlin, 1. edition, 2004.

15 Alpeshkumar K. Malde, Le Zuo, Matthew Breeze, Martin Stroet, David Poger, Pramod C. Nair, Chris Oostenbrink, and Alan E. Mark. An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0. *Journal of Chemical Theory and Computation*, 7(12):4026–4037, 2011. doi:10.1021/ct200196m.

16 Brajesh K. Rai and Gregory A. Bakken. Fast and accurate generation of ab initio quality atomic charges using nonparametric statistical regression. *Journal of Computational Chemistry*, 34(19):1661–1671, 2013. doi:10.1002/jcc.23308.

17 Matthew J. Saltzman. COIN-OR: An open-source library for optimization. In Søren S. Nielsen, editor, *Programming Languages and Systems in Computational Economics and Finance*, pages 3–32. Springer, Boston, MA, 2002. doi:10.1007/978-1-4615-1049-9_1.

18 Bradley Sherborne, Veerabahu Shanmugasundaram, Alan C. Cheng, Clara D. Christ, Renee L. DesJarlais, Jose S. Duca, Richard A. Lewis, Deborah A. Loughney, Eric S. Manas, Georgia B. McGaughey, Catherine E. Peishoff, and Herman van Vlijmen. Collaborating to improve the use of free-energy and other quantitative methods in drug discovery. *Journal of Computer-Aided Molecular Design*, 30(12):1139–1141, 2016. doi:10.1007/s10822-016-9996-y.

19 U. Chandra Singh and Peter A. Kollman. An approach to computing electrostatic charges for molecules. *Journal of Computational Chemistry*, 5(2):129–145, 1984. doi:10.1002/jcc.540050204.

20 Johannes H. Voigt, Bruno Bienfait, Shaomeng Wang, and Marc C. Nicklaus. Comparison of the NCI Open Database with Seven Large Chemical Structural Databases. *Journal of Chemical Information and Computer Sciences*, 41(3):702–712, 2001. doi:10.1021/ci000150t.