Stereoisomerism in cocrystal structure analysis

Daan Sprenkels
Radboud University Nijmegen

1st February 2016

The phenomenon of cocrystallisaton of different neutral compounds is of interest to both the academic world and pharmaceutical industry. A systematic analysis of cocrystals present in crystallographic databases can form the basis for cocrystal prediction.

A toolchain has therefore been composed to match coformer structures to cocrystal structures from the Cambridge Structural Database, while taking chirality and other isomeric features into account. The algorithm relies on the Open Babel toolkit, has reasonable performance and is able to find a correct assignment for 95 % of all cases for which two coformer candidates are proposed for one cocrystal. In the used candidates set, 976 cases (14 %) of ambiguous assignments were resolved and 843 incorrect assignments (12 %) were identified.

Besides the matching algorithm, a new method for determining a quantitative measure for molecular similarity is proposed. This method analyses the cross correlation function of histograms generated from molecules. (Reciprocal) intramolecular distance histograms have been used to test this method. A test set has been composed of similar and dissimilar structure pairs. After optimising the parameters, the routine passed 88 % of the test cases by choosing the more similar pair.

Contents

1 Introduction

1.1 Cocrystallisation and molecular similaritiy

Cocrystallisation is a method of packing different neutral molecular structures together in a non-covalent manner and has attracted increasing attention in recent years [1]. It is used for altering the properties of pharmaceutical materials [2,3], purification of enantiomeric structures [4,5] and in other fields [6].

As most experimental techniques used in cocrystal searching and engineering are labor and time expensive, a lot of effort has been put in computationally understanding and predicting cocrystallisation behaviour [6–8] using lattice energy calculations. These studies are generally applied to small predefined lists of cocrystal candidate structures and are therefore unsuitable for predicting new coformer candidates from large (database) sets. Methods based on scaffold hopping could help us solving the challenge of finding new coformer candidate structures from large structural databases, as this method was succesfully applied many times in drug design [9]: If coformer A cocrystallises with coformer B, we may look for coformers similar to B as candidates for cocrystallisation with A, especially when such indications can be found in databases.

1.2 Scope of this thesis

To mine databases with respect to cocrystallisation, we will need bulk routines which can tell which coformer structures combine to give cocrystals. The available search methods[1] are isomerism insensitive and thus produce collisions in their results when one of the coformers has multiple possible isomeric arrangements. The first topic of this thesis is therefore developing methods for matching CSD coformer structures to cocrystal structures, while taking isomerism into account.

[1] Cambridge Structural database: ConQuest, Mercury et cetera.

Apart from examining whether structures are topologically identical, a robust method for the determination of 3D-structural similarity is studied. The traditional method of scoring the similarity between different structures is based on identifying structural traits that are the same across different structures. The second topic of this thesis is describing the efforts of using a new molecular similarity index, which is based on generating artificial histograms from CSD structures and comparing them using a cross correlation function approach developed earlier in our group [10].

2 Background and theory

In chemoinformatics, *molecular identifiers* are used to respresent molecular structures directly. Examples of these identifiers are SMILES strings [11,12] and InchI strings [13]. Molecular identifiers often differ from one another in the set of features that they describe. For instance: Some may encode isomeric features, where others may not. Molecular identifiers are useful for determining molecular *equality*.

Molecular *similarity*, however, is determined using *molecular descriptors* (sometimes called molecular fingerprints). Molecular descriptors are data types that represent compounds indirectly. They are often generated from molecular identifiers, but may also be obtained using experimental methods: An NMR spectrum or powder diffraction pattern may just as well be considered a molecular descriptor.

2.1 Molecular identifiers and descriptors

2.1.1 SMILES strings

SMILES strings are character strings describing chemical structures in the widely adopted SMILES notation [11, 12]. Traditionally, SMILES strings do not encode cis-trans isomerism and chirality, but some modern SMILES generating algorithms *do* encode these properties, which characterises these strings as isomerism dependent. SMILES strings follow an easy specification, can be easily read by humans and are suitable for basic string manipulation.

A drawback of SMILES strings is, however, that individual structures can be described using a wide range of different strings, for example: Acetone can be described by both `CC(C)=O` and `CC(=O)C`. In our research we will need to unambiguously compare different structures, so traditional SMILES strings will not suffice. Canonical SMILES strings (`can` strings), as produced by Open Babel [14], will be used to make unambiguous comparison possible.

Digitally handling SMILES strings is not a hard task. One could consider using SMILES strings for more than just matching identical structures. However, the usage of SMILES strings does not seem to be very powerful when introducing the manipulation of structures. The method described in this thesis is a read-only method and performs several canonicalisation steps, which makes the usage of SMILES strings possible at all. For other problems than the one proposed in this thesis, using SMILES strings would probably not be appropriate.

2.1.2 Histogram descriptors

Histogram descriptors are histograms which can be constructed using many different types of functions and spectroscopic methods [10, 15, 16]. Because of this, a histogram descriptor may be used to describe a lot of features that other data types (like SMILES strings) lack, like intramolecular distances, fractal dimension, shape and chemical activity. On the other hand, these descriptors will always miss certain molecular features and cannot be reverted to their original 3D structures. Although these histograms are usually one dimensional, they are not limited to one dimension, which makes them very versatile for data mining practices.

2.1.3 Open Babel

Open Babel [14] is an open source toolkit used to convert chemical data formats. It effortlessly parses and generates multiple formats (like SMILES, InChI, MOL, MOL2). Babel is called from the command line[2] using one of the `{o,}babel`[3] binaries. This allows a programmer to use Open Babel to do conversions in bulk.

2.2 The Cambridge Structural Database

The Cambridge Structural Database[4] (*CSD*) is a collection of crystal structures of small molecules [17]. It allows advanced queries and bulk exporting of structures. However, some of the structures in the CSD contain imperfections, like omitted hydrogen atoms (when they are needed), unspecified disordered atoms and missing bonds. In these cases, the database will mostly contain a family of entries for distinct crystal compounds. In this research, all the structures bound to a specific refcode family (e.g. not only `ABEKUN`, but each of `ABEKUN{,01,02}`) will be used.

[2] Open Babel also has bindings for a handful of programming languages, which may be used as well.

[3] Version: Open Babel 2.3.2

[4] In this research, version v5.36 is used.

3 Methods and approach

A set of triplets consisting of nine thousand unique cocrystal structural refcodes and coformer candidates have been provided using the Cambridge Structural Database. A fragment of this list is shown in figure 1. All the structures from the set have been exported to a large chemical table file (`.sd`), with the CSD "exclude disordered atoms" option set.

```
ABEKUN   RESORA01 EVUSIX
ABEKUN   RESORA01 AZSTBB
ABEKUN   RESORA   EVUSIX
ABEKUN   RESORA02 AZSTBB
ABEKUN   RESORA02 EVUSIX
ABEKUN   RESORA03 AZSTBB
ABEKUN   RESORA13 EVUSIX
ABEKUN   RESORA13 AZSTBB
ABEKUN   RESORA03 EVUSIX
ABEKUN   RESORA   AZSTBB
```

Figure 1: triplets of cocrystals and coformers: For `ABEKUN{,01,02}` alone, there are 60 combinations of refcodes that can be used (with coformer refcode families `RESORA`x, `EVUSIX`y and `AZSTBB`z).

Two programs for examining the structures are built. The Python programming language is used in combination with the NumPy and Scipy toolkits, because it is ideal for prototypical development [18].

The first of these routines is the *SMILES toolchain*. This program checks for molecular equality, and is able to filter cocrystal candidate lists (as the one in figure 1).

The second routine parses molecular 3D structures and generates intramolecular (reciprocal) distance histograms from these structures. It is then able to determine molecular similarity coefficients using two different distance histograms.

3.1 SMILES toolchain

Using Open Babel, each chemical table structure of which 3D coordinates are present in the CSD is converted to a `can` string. Equivalently, the (reference) SMILES strings from the CSD are converted to `can` strings.

In SMILES strings, unattached structures are split by a dot character (`'.'`), which makes the extraction of separate coformer structures from these `can`

strings easy. When extracting the structures, it is assumed that singleton atom structures represent disorder, so these are not extracted. After extracting each of the coformer strings, they are reconstructed in a canonical form (using Babel) and each of the SMILES strings is simplified by ommitting all brackets and `H` symbols. If no coordinates were present in the CSD, Babel returns an empty `can` string, and triplets needing this string will be labeled `skip/nocoords` accordingly.

Each coformer candidate string should contain only one unique structure, but this is not always the case. In the coformer candidates set, there are some coformer structures that are not made up of a single molecule, but are really cocrystal structures themselves. Examples refcodes of this type of incorrectly labeled coformer structures are `SODVUC` and `SABNAK`, which are both cocrystal structures made up of two different stereoisomers of the same compound. The toolchain will filter these structures by skipping a triplet containing coformer structures, which contain more than one unique structure.[5] It will label such a triplet `skip/notonestring`.

By means of sanity check, stereospecific information is stripped from each of the string and each one is matched to their reference CSD (`can`) string. If a mismatch is encountered during this sanity check, it is assumed that a bad `can` string has been produced by Babel, which can occur for different reasons. These bad strings are filtered from the set of `can` strings and for this case the toolchain will label a triplet containing these kinds of strings `skip/invcan`. We will restore the stereospecific information, because it will be needed in the next step.

Finally, the program will check if for each cocrystal candidate string, any cocrystal string matches. If this is the case, the program will label the triplet `result/match`. It will yield `result/mismatch` otherwise.

Another defiance encountered in converting the chemical table structures to `can` strings is that Open Babel may crash or leak memory indefinitely (until killed by the kernel) while converting structures with certain types of disordered patterns (examples of these kinds of structures are `MUBJAU` and `OBEQUH`). We will use the `setrlimit(2)` system call to limit the process' resources to prevent dangerous process behaviour. The amount of structures that causes these kinds of difficulties is, even for huge data sets, insignificant. Although the results will not be fully complete, it is not expected that the loss of this handful of structures will impact the results in any way when

[5]Some structures in the CSD contain unattached atoms, often to visualise the disordered state of those atoms. In SMILES strings, these atoms are considered separate structures, but are not needed for the SMILES toolkit. Accordingly, these unattached atoms will be excluded from each `can` string just before performing this step of filtering coformer structures containing more than one unique structure.

dealing with large screenings. Nonetheless, triplets containing any of these refcodes will be labeled `skip/panic`.

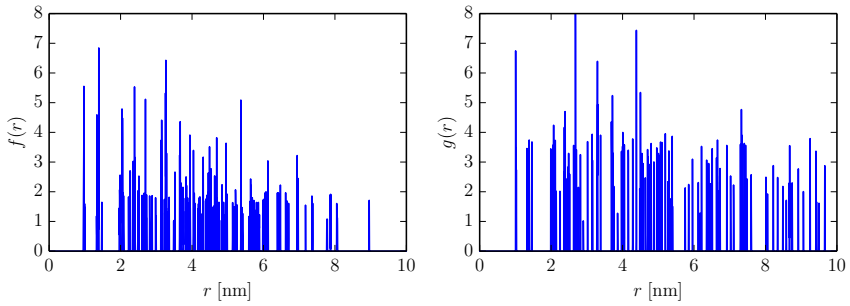
3.2 Histogram based similarity

3.2.1 Intramolecular distance

The histogram type $H(r)$ that is chosen to fulfill the role of prototype molecular descriptor is one based on intramolecular distances. It is defined in equation 1. Two example histograms (for `EVUSIX` and `AZSTBB`) are provided in figure 2.

$$H(r) = \sum_{i \neq j} C_{ij} \delta_{r d_{ij}} \tag{1}$$

The histogram function $H(r)$ is defined for a range from 0 to r_{max}; $[r] = \text{nm}$. Each interatomic distance in the molecule $d_{ij} = |\vec{r}_j - \vec{r}_i|$, where $i \neq j$, is added at $r = d$ with a value of C_{ij}. C_{ij} is a constant based on the properties of atoms i and j, which will take a value of 1 as yet. The histogram does not have to be normalised, because the final S values will be normalised in the end.



(a) `EVUSIX` (b) `AZSTBB`

Figure 2: example intramolecular distance histograms $f(r)$ and $g(r)$ of which the similarity will be determined

Double iteration will be needed for the generation of these histograms, so a complexity of $\mathcal{O}(n^2)$ is expected for this algorithm, where n is the amount of atoms in the molecule. Therefore, the performance of this routine will degrade drastically with the amount of atoms in a structure. This method would thus be less suitable for larger molecules (like proteins), considering calculation time.

3.2.2 Calculating the similarity coefficient S

The similarity of two histograms can be determined using different methods. In this thesis, the method of calculating the similarity of different histograms will be the method described in [10]:

Provided two histograms $f(r)$ and $g(r)$, the cross correlation function $c_{fg} = f \star g$ and the autocorrelation functions $c_{ff} = f \star f$ and $c_{gg} = g \star g$ are defined. A triangle weighting function $w(r)$ is defined in equation 2 and visualised in figure 3 in green. In this definition l will be the width of the triangle in $w(r)$.

$$w(r) = \begin{cases} 1 - |r|/l & \text{if } |r| < l \\ 0 & \text{if } |r| \geq l \end{cases} \tag{2}$$

Each of the auto and cross correlation functions is multiplied by $w(r)$. For c_{fg}, this results in the *modified* or *filtered* cross correlation function, shown in 4 in blue. The similarity coefficient S will be the integral of the modified cross correlation function c_{fg}, normalised using the integrals of the modified autocorrelation functions, as described by equation 3. Because S is normalised, the similarity coefficient will always be between 0 and 1.

$$S_{fg} = \int w(r) c_{fg}(r) \mathrm{dr} \Big/ \left(\int w(r) c_{ff}(r) \mathrm{dr} \int w(r) c_{gg}(r) \mathrm{dr} \right)^{\frac{1}{2}} \tag{3}$$

When peaks in a pair of histograms are found at similar r values, the cross correlation function $c_{fg}(r)$ is large near $r = 0$. In the case that peaks are not at similar positions, $c_{fg}(r)$ will be larger around higher values of r. Because of the multiplication of $c_{fg}(r)$ with $w(r)$, the former case will yield a higher similarity value than the latter case.

An algorithm implementing this would have to doubly iterate over the points in each histogram. The complexity of this algorithm will be at least $\mathcal{O}(k^2)$, where k is the amount of points in each spectrum. The complexity of the algorithm does not depend on any structural traits, which makes this method versatile for different types of structures, but note that complexity of the algorithms generating histograms will presumably depend on structural traits.

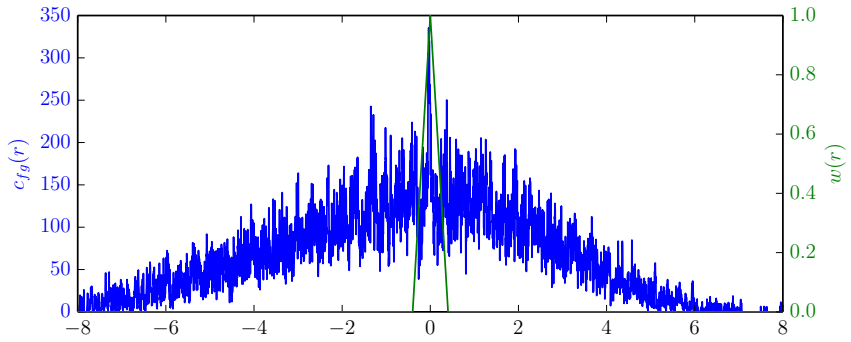


Figure 3: correlation histogram of the histograms shown in 2 in blue; triangle weighting function $c_{fg}(r)$ visualised in green.

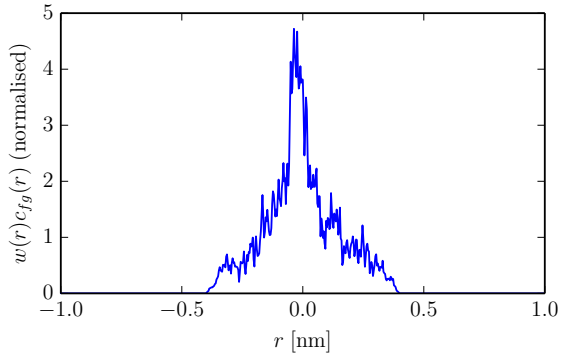


Figure 4: correlation function after multiplying with $w(r)$ and normalising using c_{ff} and c_{gg}.

It is expected that the quality of the distance histograms will vary because of different factors, for instance: One can expect bad results when examining large molecules, or when examining molecules with a high degree of conformational freedom. In these cases the conformational variations may often supersede the triangle width of the weighting function (equation 2), which will have a bad impact on the final S value.

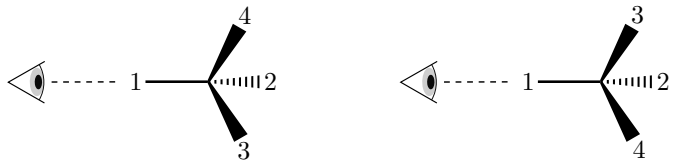
We can, however, introduce some minor optimisations, and validate them using the results obtained from the SMILES procedure (this is described in section 3.2.4).

- Fill in atomic constant C_{ij}, so that the appearance of certain types of atoms will impact the histograms in a different manner.
- Exclude hydrogen atoms from each structure, before generating a histogram (equivalent to deciding that $C_{ij} = 0$ when i or j is a hydrogen atom).
- Use a reciprocal distance scale, instead of an absolute one. This would result in smaller intramolecular distances weighing more than large distances. As a result, conformational freedom would have less impact on the histogram variations.

3.2.3 Chirality index

A shortcoming of these distance histograms is that they are not chirality dependent. To expand the histogram descriptor we will introduce a chirality index, which will be determined apart from the histograms, so that the chirality of structures can be examined as well.

Molecular chirality spans a wide range of different types [19]. The following proposed chirality index will only describe the chirality around single atoms with four distinguishable groups attached, as this is the most common type of chirality. Stated is that the descriptor should describe nothing else than molecular chirality.



(a) positive chirality (b) negative chirality

Figure 5: Examples of AWCS chirality classification. The numbers on the atoms are (example) AWCS values. In the left figure (5a), viewing from the position of the eye, the AWCS values are ordered in counterclockwise direction (positive atom). In the right figure (5b), the values are ordered in clockwise direction (negative atom).

First the k-ordered *atomic walk counts sum* topological index value [20] is obtained for each node in the hydrogen-excluded molecular graph, where k is one less than the number of nodes in the graph. Using these AWCS

values, each atom connected to four different groups will be identified. For each of these chiral atoms, the bonded atoms will be ordered by AWCS value (hydrogen atoms will be assigned an AWCS value of 0). The bonded atom with the lowest AWCS value will be chosen as offset 1. Offset 2 will be the atom with the next-to-lowest AWCS value. Finally, when viewing the center from offset 1 and disregarding the ACWS value of offset 2:

- if the remaining AWCS values are ordered in *counter-clockwise* direction, the chiral atom will be called a *positive* chiral atom;
- if the remaining AWCS values are ordered in *clockwise* direction, the chiral atom will be called a *negative* chiral atom.

An example of this mechanism is provided in figure 5. The chirality index will be made up of an unordered list of assignments for all the chiral atom (indexed by their AWCS values) in the structure. The results will be used for filtering the test set described in the next section.

3.2.4 Validating S

561 test cases have been composed using the results from the SMILES method. Each test case will be made up of two triplets from the original set. Of these two triplets only one consists of a cocrystal structure and its correct coformer structures. The other triplet will have one coformer differing in stereoisomeric structure, which does not occur in the cocrystal structure.

Because the distance histograms will not be sensitive for all types of stereoisomerism, the test cases that are based on stereoisomeric differences are filtered from the test set using the method described in 3.2.3. This filter method does not exclude *all* stereoisomeric cases, for example: Cases of axial chirality are not identified. However, no cases of these types of chirality are present in the test set. Furthermore, cases of diastereomeric isomerism are also not excluded, as differences between these structures should show through the distance histograms. After filtering the test cases, 230 usable test cases remain.

For each triplet in the filtered test set, the differing coformer structures are compared to each separate structure in the cocrystal chemical table. For each differing coformer a list of S values is obtained. It is assumed that, for each coformer, the largest of the S values is associated to the correct coformer structure. The two largest similarity values for each coformer are compared. If $S_{\text{correct,max}} > S_{\text{other,max}}$, the test case is considered *passed*. If

the similarity value obtained from the wrong coformer is not less than the similarity from the correct coformer, the test is considered *failed*.

Because the distance histogram is a shallow descriptor by design, it is not expected that the results will be exceptionally accurate. It *is* expected that the test set will show that the use of these histogram based descriptors can be used to identify distinct structures, and thus be used as a measure of similarity. If this is the case, we could look into more complex histogram descriptors, as to increase the performance of the method.

4 Results

4.1 Coformer matching using SMILES strings

The SMILES toolchain has been applied to the cocrystal candidates set. The triplets that had refcodes of compounds of which no coordinates were available in the CSD (those labeled `skip/nocoords`) have been excluded from the results. The ratio of triplets that yielded a (mis)match to the ones that could not be processed is shown in table 1. The table provides data for each individual triplet in the set and for each grouped combination of refcode families (i.e. triplets (`ABEKUN`, `RESORA`, `AZSTBB`) and (`ABEKUN01`, `RESORA`, `AZSTBB`) are grouped, and counted only once). Applying the whole toolchain to a data set of a small half million of triplets, containing 17 thousand different structures, took about 40 minutes[6].

Table 1: results of the SMILES comparison method (statistics of entries with refcodes lacking coordinates are excluded); a triplet is considered *unsimilar* if the combination of cocrystal and the coformers families is unique

	individual triplets		unsimilar triplets	
`result/*`	283538	95,45 %	7776	94,75 %
`skip/*`	13516	4,55 %	431	5,25 %
Total	297054	100 %	8207	100 %

The toolchain was able to determine if both coformers are present about 95 % of the time, when looking at the statistics of individual triplets as well as grouped triplets. Of the 5 % of triplets that were skipped, the statistics of the reason they were skipped are provided in table 2.

Table 2: distributions of the skipped entries (statistics of entries with refcodes lacking coordinates are excluded); a triplet is considered *unsimilar* if the combination of cocrystal and the coformers families is unique

	individual triplets		unsimilar triplets	
`skip/invcan`	11092	82,07 %	331	76,80 %
`skip/notonestring`	2386	17,65 %	96	22,27 %
`skip/panic`	38	0,28 %	4	0,93 %

The majority of the time, `can` strings failing the `skip/invcan` sanity check failed, because in the CSD these structures were visualised with disorders. These structures are not "real" (valid) structures, but shown in a manner, so that fellow researchers are able to interpret the way atoms are disordered in the crystal. This makes these specific cases impossible to convert to `can`

[6]Measured on a single `Intel(R) Xeon(R) CPU E5-2430L v2 @ 2.40GHz` core.

strings using Babel and thus impossible to match to other structures. This is, however, only the case in a small amount of triplets (less than five percent of any produced data set), so it does not have a large impact on the total results.

The amount of triplets that had compounds which contained more than one structure (`skip/notonestring`) was less than one percent. This is expected, as these kinds of compounds do not occur very often in the CSD.

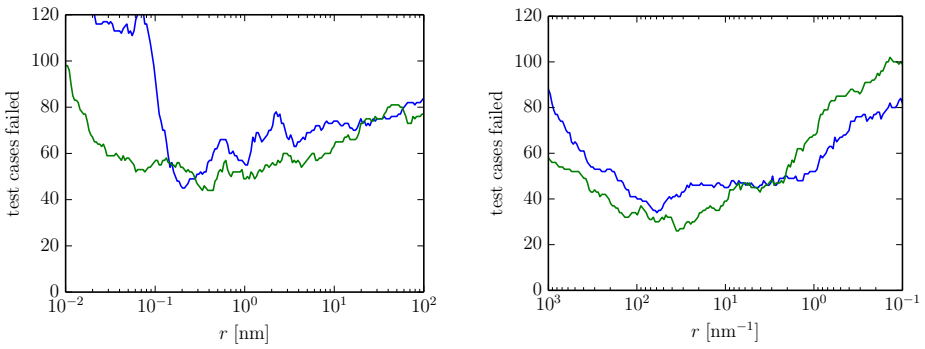
Finally, the amount of `skip/panic` cases is negligible. The only refcodes that have triggered this error are `MUBJAU` and `OBEQUH`.

From the original set 976 ambiguous assignments were cleared up (14 %). Furthermore, 843 triplets have been identified where the original assignment was *incorrect* (and of which a correct assignment did not occur in the set, another 12 %).

4.2 Intramolecular distance based similarity

Using the test set, a suitable l (triangle width) value is found by minimising the amount of failing test cases for given values of l. The results are shown in figure 6. When the variation of excluding hydrogen atoms is introduced, the routine seems to pass a significant amount of extra test cases for most values of l.

The usage of reciprocal distances, instead of absolute distances, was introduced with a newly calculated reciprocal value l (≈ 0.028 nm^{-1}). In both the cases measures were done where hydrogen atoms were kept and excluded.



(a) absolute distance

(b) reciprocal distance

Figure 6: amount of failed test cases per triangle width value of $w(r)$ (lower is better). Hydrogen atoms included: blue. Hydrogen atoms excluded: green

With the variation of using reciprocal distances, the amount of passed tests increased a lot. With this improvement, the minimum amount of failed tests was 11 % instead of 19 %.

The best fail rate that was acquired was 11 % of the test cases, while excluding hydrogen atoms, using reciprocal distances and using a value $l = 0.0281$ nm^{-1}.

5 Discussion

The SMILES identification method works well. Before applying the sanity check of matching the produced `can` strings to the SMILES strings exported from the CSD, some false negatives would be found when taking samples of the result set by hand. After implementing the sanity check where the produced `can` strings are first matched against the CSD SMILES strings, no false positives and no false negatives were found anymore. This method could thus be a basis for building a general cocrystal index.

The performance of this method seems to be better than expected. Although performance was not a critical concern, the toolchain completed within the hour, on a slightly-above-average CPU. The toolchain is also very easy to parallelise, which could overcome long calculation times when the cocrystal candidates set gets larger.

The method of basing similarity on distance histograms seems promising. Considering the small amount of optimising variations, a final fail rate of 11 % looks to be an acceptable result, but the method still has some drawbacks:

- The correlation between several properties of the tested molecules and their absolute similarity is not yet thoroughly investigated in any kind. It could be very possible that larger molecules will *just* produce smaller S values, which would make this method unreliable when calculating the similarity between molecules of different sizes.
- The test cases allowed to compare different similarities to each other. It was tested that, given two different structures of which one was known to be *more similar than the other*, sound conclusions could be produced. However, the individual S values diverted from 0.5 to 1.0, in the case of both similar and dissimilar structure pairs. There currently exists no "neutral similarity value" S_0. The lack of this offset makes it impossible to determine an absolute similarity for structures. The produced values of S can only be used to order and compare different structural similarities.

6 Conclusion

An algorithm for finding exact matches of cocrystal structures from the CSD, using canonical SMILES strings, has been built. The algorithm relies on the Open Babel toolkit and has good performance. The part of the structures that can be processed is 95 %. While it is not possible to examine all the structures from the CSD, the cases of which the program is not able to tell if it contains a match, involve only exceptional structures and the amount of these cases is small.

The method for determining structural similarity, using intramolecular distance histograms, is promising and could be generally applicable in molecular similarity issues. Due to the prototypic nature of this method, it is not yet reliable. For confirming the effectiveness of this method, this method should be expanded on by trying different (combinations of) descriptor formats and similarity calculation functions.

7 Further research and concluding remarks

7.1 On structural examination

While researching the possibilities of easily comparing different structures, it occurred that tools for comparing the structures, while taking basic isomerism into account, were sparsely available. This was even the case when identical comparison was the only task. Tools for examining the isomerism itself seemed to be sparse as well. Molecular descriptors like MCDL [21, 22] and PZIM [23] are independent of more complex isomerism features, like cis-trans isomerism over rings. When these features are crucial, one will have to fall back to using descriptors like InChI strings. But because of the relative complex nature of InChI strings, it may be preferable to build more low level structural examination routines. Using these routines, Open Babel may eventually be eliminated as dependency for the SMILES examination toolchain. In section 3.2.3, a method of examining molecular chirality using atomic walk counts sums is described. To add to this method, we propose a conceptual method to expand our ability to examine the cis-trans isomerism of structures, described in appendix 1, which could easily be implemented.

At the moment, the algorithm that was composed is able to find matches (or mismatches) for cocrystals with just two different coformer ligands. For handling cocrystal structures with more than two different coformers, this method will have to be expanded.

7.2 On distance histograms

The method of using distance histograms for calculating the similarity seems to work, but can not be called robust or reliable at the time.

A current problem, as described in section 4.2, is the lack of a similarity offset. The method could be also expanded so that the resulting S values are normalised relative to one another.

Only one similarity calculation method has been used. While this calculation method may perform well in [10], other methods are widely available [24, 25] and may perform better. This research gives us an opportunity to verify the performance of the method described in [10], but it would be wise to perform a comparison of different similarity calculation methods nonetheless.

Only intramolecular distance histograms have been used for this method. Other structural characteristics have not been used, but some could presum-

ably contribute to the molecular description used to determine the similarity. Examples of these kinds of features could be:

- molecular size,
- atomic numbers, masses, radii, periods (in periodic table) and groups,
- electrostatic or Van der Waals potential maps, and
- fractal dimension [15].

Note that, for the sake of simplicity, only one-dimensional histograms and correlation histograms have been used. Histogram descriptors do not require to be one-dimensional at all, as the similarity calculation method that is used can easily be extended to handle multidimensional histograms. This gives us the opportunity to expand to using multidimensional data structures as descriptors and using multivariate regression methods for the calculation of S.

8 References

[1] S. Childs, "The Reemergence of Cocrystals: The Crystal Clear Writing Is on the Wall," *Cryst. Growth Des*, vol. 10, no. 9, pp. 4208–4211, 2009.

[2] N. Schultheiss and A. Newman, "Pharmaceutical Cocrystals and Their Physicochemical Properties," *Cryst. Growth Des.*, vol. 6, no. 9, pp. 2950–2967, 2009.

[3] T. Friscic and W. Jones, "Benefits of cocrystallisation in pharmaceutical materials science: an update," *J. Pharm. Pharmacol*, vol. 62, no. 11, pp. 1547–1559.

[4] M. Habgood, "Analysis of Enantiospecific and Diastereomeric Cocrystal Systems by Crystal Structure Prediction," *Cryst. Growth Des.*, vol. 10, no. 13, pp. 4549–4558, 2013.

[5] M. Eddleston, M. Arhangelskis, T. Friscic, and W. Jones, "Solid state grinding as a tool to aid enantiomeric resolution by cocrystallisation," *Chem. Commun.*, vol. 48, pp. 11340–11342, 2012.

[6] J. Zhou, M. Chen, W. Chen, L. Shi, C. Zhang, and H. Li, "Virtual screening of cocrystal formers for CL-20," *J. Mol. Struc*, vol. 1072, pp. 179–186, 2014.

[7] N. Issa, P. Karamertzanis, G. Welch, and S. Price, "Can the Formation of Pharmaceutical Cocrystals Be Computationally Predicted? I. Comparison of Lattice Energies," *Cryst. Growth Des.*, vol. 9, no. 1, pp. 442–453, 2009.

[8] P. Karamertzanis, A. Kazantsev, N. Issa, G. Welch, C. Adjiman, C. Pantelides, and S. Price, "Can the Formation of Pharmaceutical Cocrystals Be Computationally Predicted? 2. Crystal Structure Prediction," *J. Chem. Theory Comput.*, vol. 5, no. 5, pp. 1432–1448, 2009.

[9] M. S. H. Böhm, A. Flohr, "Scaffold hopping," *Drug Discovery Today: Technologies*, vol. 1, no. 3, pp. 217–224, 2004.

[10] R. de Gelder, R. Wehrens, and J. Hageman, "A generalized expression for the similarity of spectra: Application to powder diffraction pattern classification," *J. Comput. Chem.*, vol. 22, no. 3, pp. 273–289, 2001.

[11] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *J. Chem. Inf. Model.*, vol. 28, no. 1, pp. 31–36, 1988.

[12] D. Weininger, A. Weininger, and J. L. Weininger, "SMILES. 2. Algorithm for generation of unique SMILES notation," *J. Chem. Inf. Model.*, vol. 29, no. 2, pp. 97–101, 1989.

[13] S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi, and I. Pletnev, "InChI - the worldwide chemical structure identifier standard," *J. Cheminform.*, vol. 5, 2013.

[14] N. O'Boyle, M. Banck, C. James, C. Morley, T. Vandermeersch, and G. Hutchison, "Open Babel: An open chemical toolbox," *J. Cheminform.*, vol. 3, no. 1, p. 33, 2011.

[15] V. Grigor'ev and O. Raevskii, "Fractal dimension of the interatomic distance histogram: New 3D descriptor of molecular structure," *Russian Journal of General Chemistry*, vol. 81, no. 3, pp. 449–455, 2011.

[16] S. Gu, P. Koehl, J. Hass, and N. Amenta, "Surface-histogram: A new shape descriptor for protein-protein docking," *Proteins: Structure, Function, and Bioinformatics*, vol. 80, no. 1, pp. 221–238, 2012.

[17] F. H. Allen, "The Cambridge Structural Database: a quarter of a million crystal structures and rising," *Acta Crystallographica Section B*, vol. 58, pp. 380–388, Jun 2002.

[18] D. Garey and S. Lang, "High Performance Development with Python." `www.scientificcomputing.com/articles/2008/11/high-performance-development-python`, 2008. Accessed: 2015 – 11 – 16.

[19] A. Golbraikh, D. Bonchev, and A. Tropsha, "Novel Chirality Descriptors Derived from Molecular Topology," *J. Chem. Inf. Model.*, vol. 41, no. 1, pp. 147–158, 2001.

[20] G. Rücker and C. Rücker, "Counts of all walks as atomic and molecular descriptors," *J. Chem. Inf. Model.*, vol. 33, no. 5, pp. 683–695, 1993.

[21] A. A. Gakh and M. N. Burnett, "Modular Chemical Descriptor Language (MCDL): Composition, Connectivity, and Supplementary Modules," *J. Chem. Inf. Comput. Sci.*, vol. 41, no. 6, pp. 1494–1499, 2001.

[22] A. A. Gakh, M. N. Burnett, S. V. Trepalin, and A. V. Yarkov, "Modular Chemical Descriptor Language (MCDL): Stereochemical modules," *J. Cheminform.*, vol. 3, 2011.

[23] A. E. Berglund and R. D. Head, "PZIM: A Method for Similarity Searching Using Atom Environments and 2D Alignment," *J. Chem. Inf. Model.*, vol. 50, no. 10, pp. 1790–1795, 2010.

[24] Y. Cao and L. Petzold, "Accuracy limitations and the measurement of errors in the stochastic simulation of chemically reacting systems," *Journal of Computational Physics*, vol. 212, no. 1, pp. 6 – 24, 2006.

[25] NIST/SEMATECH, "e-Handbook of Statistical Methods." `http://www.itl.nist.gov/div898/handbook/`. Date: 2013 – 10 – 30.

[26] International Union of Pure and Applied Chemistry, "IUPAC Compendium of Chemical Terminology – The Gold Book," 2009.

9 Appendix 1: A method for classifying cis-trans traits over double bonds and ring structures

These methods were thought of with the development of the chirality examination method (section 3.2.3). It was never implemented, because, during my research, it was never needed to filter structures based on cis-trans isomerism. When implemented, both of the following methods could provide us with a lot of bulk information on isomeric differences between different molecules.

A Double bonds

Distinguishing whether double bonds are cis or trans boils down to applying the method described in section 3.2.3. Consider the double bond described in figure 7. For each of the atoms the atoms attached to the doubly bonded atoms (atoms $R_1 - R_4$), we will determine their hydrogen excluded ACWS value (just as in section 3.2.3 we will define hydrogen ligands to have an AWCS value of 0).

We will define group R_1 to be the group with the lowest AWCS value. We will then decide that of the attached atoms on the other side of the bond, we will again select the one with the lowest AWCS as offset. If these two atoms are cis-configured, decide that the bond is *cis*, and if they are trans-configured, decide that it is *trans*.

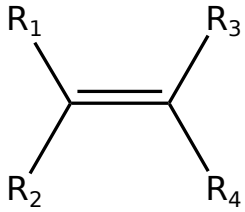


Figure 7: double bond

In the case of figure 7 this is (when using the atom's AWCS values for ordering) analogous to the following:

$$\text{a double bond is } \begin{cases} \textit{cis} & \text{if } R_3 < R_4 \\ \textit{trans} & \text{if } R_3 > R_4 \end{cases} \tag{4}$$

Note that these definitions of cis and trans do not comply with the "real" (IUPAC Gold Book) definition [26], as we are using AWCS values as a

canonicalisation method for the attached groups, instead of the conventional cis-trans priority rules.

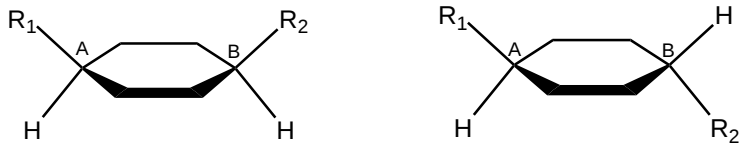
B Ring structures

First the following symbols are defined:

- M: Molecular graph, contains n nodes.
- A, B: Molecular centers (atoms) to be examined.
- A_i, B_i: Atoms attached to A and B, described by the $(n-1)$-order of AWCS values of the attached atoms; this order will be described further below.
- $A_i = B_j$, when A_i and B_j are considered *equivalent*; their ACWS values are equal.
- $A = B$: Centers A and B are both positive or both negative; refer to section 3.2.3 or figure 5 for the meaning of *positive* and *negative*.

The current algorithm algorithm (section 3.2.3) will classify every sp^3 center with four different groups attached. So we will propose an algorithm that can handle sp^3 centers with two groups having identical AWCS values.

Consider the molecule M, viewed in figure 8.



(a) cis configuration (b) trans configuration

Figure 8: example molecule containing a cis-trans ring

For every A_i, A_j where $A_i = A_j$, find the following shortest path:

$$P = (A, A_i, \ldots, B, \ldots, A_j, A)$$

with the constraints that:

1. Every vertex (bond) is walked along not more than once.
2. The path is, expressed in AWCS values, a palindrome.

Using P, we will define A_1 to be A_i and A_2 to be A_j. B_1 and B_2 will be defined so that the following holds:

$$P = (A, A_1, \ldots, B_2, B, B_1, \ldots, A_2, A)$$

A_3, A_4 and B_3, B_4 will be defined in a canonical manner. For example, by order of AWCS value. We will now be able to distinguish different cis-trans isomers by testing if $A = B$ or if $A \neq B$.

In complex cases, when multiple rings are traversed, this method will fail. These cases can be identified by the fact that, in the path-finding step, multiple equal shortest paths will be found. An example of this is axial chirality, as shown in figure 9.

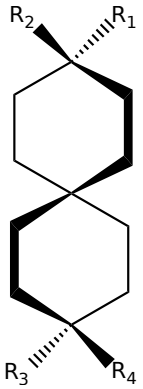


Figure 9: example case of axial chirality. None of the methods described in this thesis are able to classify the chirality of this molecule, as the two rings together make up a single (tetrahedral) chiral center.

This shows the need for more routines that are able to identify different types of isomerism. This fact can also be interpreted as a sign that this method is long from perfect and more research should be done in the development of more complete algorithms, as the problem of describing the "real"[7] isomerism of molecules turns out to be more complex than anyone would intuitively expect.

[7]Here is meant: A descriptor that is both *complete* (all isomerism features are described by the descriptor) and *canonical* (there exists only *one* valid descriptor to describe the isomerism of any structure).