applied sciences

MDPI

Article

Beyond the Arbitrariness of Drug-Likeness Rules: Rough Set Theory and Decision Rules in the Service of Drug Design

Grzegorz Miebs [1], Adam Mielniczuk [1], Miłosz Kadziński [1,*] and Rafał A. Bachorz [1,2]

[1] Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznań, Poland; grzegorz.miebs@cs.put.poznan.pl (G.M.); adam.mielniczuk@student.put.poznan.pl (A.M.); rafal.bachorz@cs.put.poznan.pl (R.A.B.)
[2] Institute of Medical Biology, Polish Academy of Sciences, Lodowa 106, 93-232 Łódź, Poland
* Correspondence: milosz.kadzinski@cs.put.poznan.pl

Abstract: Lipinski's Rule of Five and Ghose filter are empirical guidelines for evaluating the drug-likeness of a compound, suggesting that orally active drugs typically fall within specific ranges for molecular descriptors such as hydrogen bond donors and acceptors, weight, and lipophilicity. We revisit these practices and offer a more analytical perspective using the Dominance-based Rough Set Approach (DRSA). By analyzing representative samples of drug and non-drug molecules and focusing on the same molecular descriptors, we derived decision rules capable of distinguishing between these two classes systematically and reproducibly. This way, we reduced human bias and enabled efficient knowledge extraction from available data. The performance of the DRSA model was rigorously validated against traditional rules and available machine learning (ML) approaches, showing a significant improvement over empirical rules while achieving comparable predictive accuracy to more complex ML methods. Our rules remain simple and interpretable while being characterized by high sensitivity and specificity.

Keywords: drug-likeness; multiple criteria decision analysis; drug design; biological activity; dominance-based rough set approach; decision rules

Citation: Miebs, G.; Mielniczuk, A.; Kadziński, M.; Bachorz, R.A. Beyond the Arbitrariness of Drug-Likeness Rules: Rough Set Theory and Decision Rules in the Service of Drug Design. *Appl. Sci.* **2024**, *14*, 9966. https://doi.org/10.3390/app14219966

Academic Editor: Alexander N. Pisarchik

Received: 20 September 2024
Revised: 27 October 2024
Accepted: 29 October 2024
Published: 31 October 2024

1. Introduction

Almost 30 years have passed since the discovery of the well-known Lipinski Rule of Five [1,2]. It allows for quick resolution if a particular compound is likely to pose expected solubility and tissue permeability properties, which are essential in drug distribution. The main idea behind the Rule of Five (RO5) is to describe the compound with four simple molecular descriptors, namely the number of hydrogen bond (HB) donors, number of HB acceptors, molecular weight, and octanol–water partition coefficient (log P). Based on experimental data, scientific intuition, and simple math, Lipinski et al. formulated approximate, high-level binary descriptors as simple inequalities capturing expected druggability properties. In the classical form, the RO5 says that a compound is likely to have poor solubility and permeability properties if the following are met:

- It contains more than five HB donors.
- It contains more than 10 HB acceptors.
- Its molecular weight > 500.
- Its octanol–water coefficient $\log P > 5.0$ (or Moriguchi $M\log P > 4.15$).

Lipinski intended to predict whether the investigated compound will likely possess physiochemical properties, allowing for its successful distribution in the human circulatory system. From a practical perspective, it gives medicinal chemists an easily applicable tool capable of providing the first information about a compound's druggability. The simplicity of utilized molecular descriptors makes an overall computational cost low and thus is

suitable for large-scale virtual screening procedures. In particular, the least promising candidates can be excluded before clinical trials.

Besides the Lipinski et al. contribution, nowadays, we know more than a dozen other similar rules. They involve various molecular descriptors, are based on slightly different assumptions, and are trained on varying data sets. The Ghose filter is an excellent example of another, well-established druggability index [3]. It also involves four molecular descriptors, but this time, they are water–octanol partition coefficient, molar refractivity, molecular weight, and the number of non-hydrogen atoms. Ghose et al. used the Comprehensive Medicinal Chemistry (CMC) database and found specific qualifying ranges of mentioned physiochemical properties covering at least 80% of the database compounds. They created several filter variants, depending on the drug class considered, e.g., anti-inflammatory, anti-depressant, or anti-hypertensive. In the literature, one can also find various drug-like scores [4–11], lead-like scores [12,13], the classification system BDDCS [14], and the exciting idea of Quantitative Estimate of Drug-Likeness (QED) [15]. Other efforts focused on the description the drug-likeness of molecules cover the application of Neural Network (NN) [16–20], Recursive Partitioning (RP) [21,22], and Support Vector Machine (SVM) [17,23].

The primary aims of this paper are three-fold. First, we construct an original data set comprising a few thousand drug and non-drug compounds. Second, we demonstrate the applicability of a suitably adapted variant of DRSA to structure the data set and automatically infer simple, reproducible, and data-driven decision rules, reflecting druggability conditions or lack thereof. Third, we rigorously validate their predictive capabilities—captured by five performance metrics—against selected empirical rules and various DRSA approaches.

2. Materials and Methods

This work considers a carefully created data set containing drug and non-drug molecules. On the one hand, we have extracted all approved drugs from the most recent version of the ChEMBL database (version 3.3) [24]. After removing the proteins, oligosaccharides, oligonucleotides, and molecules belonging to the class "unknown" and applying the standardization procedure on a SMILES representation of molecules, we ended up with 2197 compounds reflecting the so-far registered small-molecule drugs. On the other hand, the non-drug component was derived from the ZINC-22 database containing ca. 37 billion commercially available, enumerated, and searchable compounds [25]. To ensure this component comprises the same number of molecules as its drug counterpart, we sampled ZINC-22 uniformly. Since such a choice is non-unique, we repeated the sampling three times to estimate the influence of the composition of non-drug components on the quality of the resulting models. The final data sets comprised 4394 compounds containing both drug and non-drug components. They were subsequently randomly split into training and test sets, where the size of the latter component was set to 20%. The construction of the data sets was conducted while bearing in mind the study's central aim, which was to compare the existing decision analysis and machine learning approaches with the classical procedures. Therefore, we intentionally limited the training data set while leaving only small molecules and considering the same descriptor space as in the RO5 and the Ghose filter.

The data sets were subsequently structured using DRSA. It is an extension of the seminal work of Pawlak [26], handling the inconsistencies typical in multi-criteria decision-aiding processes [27]. Then, the approach leads to rules of the type *"if..., then..."*. They capture the conditions (i.e., values of one or more attributes) capable of distinguishing molecules representing two classes, drug and non-drug molecules, hence resembling the classical druggability rules. Therefore, they can be considered their natural extension obtained in a systematic, objective, and efficient way. This algorithm has been widely applied in domains such as finance [28], urban development [29], sustainability [30], energy [31],

portfolio optimization [32], and risk assessment [33]. Due to the lack of strict monotonicity of the data, we used a variant of DRSA supporting local monotonicity changes [34].

To obtain the rules, we first calculated the appropriate molecular descriptors. Here, we applied the Mordred [35] and RDkit [36] libraries. We aimed to separately consider the performance of the DRSA model against the Lipinski RO5 and Ghose filter. To achieve that, we used suitable sets of molecular descriptors in accordance with both reference approaches. For the comparison with Lipinski RO5, we provided the data spanned by the following features: molecular weight (MWt), number of HB donors (nHdon), number of HB acceptors (nHacc), and octanol–water partition coefficient (log P). To confront the performance of the DRSA model with the Ghose procedure, we utilized the data set containing molecular weight (MWt), molar refractivity (MR), number of non-hydrogen atoms (nAtoms), and again octanol–water partition coefficient (log P). In both cases, the DRSA rules were inferred with the DOMLEM algorithm [37] and then used to classify molecules from the test set [38]. The following subsection provides a detailed description of the applied method.

2.1. Dominance-Based Rough Set Approach

The decision problem in the DRSA framework is represented by a matrix where the rows correspond to a set of evaluated alternatives, $A = \{a_1, a_2, \ldots, a_n\}$, and the columns represent the evaluation criteria $G = \{g_1, g_2, \ldots, g_m\}$ along with the decision attribute D. The evaluation of each alternative $a_i \in A$ on criterion $g_j \in G$ is denoted by $g_j(a_i)$. Alternatives are split into a reference set $A^R \subset A$ used in a training phase and a non-reference set $A^{NR} \subset A$ for which the value of a decision attribute D has to be determined. Each criterion shall be defined as either cost or gain type. When a given criterion is neither, but the local monotonicity holds, it can be duplicated [34] with one copy treated as a cost component and the other as a gain counterpart. In what follows, we present the steps of DRSA when considering the specificity of the tackled problem. It involves binary classification and non-monotonic criteria.

2.1.1. Lower and Upper Approximations of Class Unions

The initial step involves data structuring that allows for identifying potential inconsistencies. For this purpose, the lower and upper approximations of class unions are established. In the case of binary classification with Cl_0 (negative class) and Cl_1 (positive class), we can directly consider decision classes instead of unions. The lower approximation $\underline{P}$ includes all alternatives that definitively belong to a specific class. When all criteria are potentially non-monotonic, $\underline{P}(Cl_r)$ contains all alternatives in Cl_r that are not indifferent to any alternative from Cl_{1-r}, i.e., $\underline{P}(Cl_r) = \{a_i \in A^R : \nexists a_k \in Cl_{1-r}, \forall g_j \in G : g_j(a_i) = g_j(a_k)\}$. In turn, the upper approximation $\overline{P}$ contains alternatives that possibly belong to the class. In our setting, $\overline{P}(Cl_r)$ contains all alternatives that are indifferent to any alternative from Cl_r, i.e., $\overline{P}(Cl_r) = \{a_i \in A^R : \exists a_k \in Cl_r, \forall g_j \in G : g_j(a_i) = g_j(a_k)\}$. Hence, if a pair of alternatives shared the same performances on all criteria while being assigned to different classes, they would belong only to an upper approximation while not being contained in the lower one.

To explain the structuring step of DRSA, we will refer to the decision matrix from Table 1. Three of five alternatives (a_1, a_2, and a_3) are assigned to class Cl_0, and the remaining two are assigned to class Cl_1. Let us first focus on the lower and upper approximations of class Cl_0. The performances of a_1 on all criteria match those of a_4, which is assigned to Cl_1. Therefore, a_1 cannot be included in $\underline{P}(Cl_0)$. However, both alternatives belong to $\overline{P}(Cl_0)$. Next, a_1 has the same performances as a_2, but since both of them are assigned to Cl_0, there is no inconsistency, and the alternatives belong to $\underline{P}(Cl_0)$ and $\overline{P}(Cl_0)$. Overall, $\underline{P}(Cl_0)$ is equal to $\{a_2, a_3\}$, whereas $\overline{P}(Cl_0)$ is $\{a_1, a_2, a_3, a_4\}$. Analogously, the lower and upper approximations of Cl_1 are $\underline{P}(Cl_1) = \{a_5\}$ and $\overline{P}(Cl_1) = \{a_1, a_4, a_5\}$.

Table 1. Example decision matrix used to illustrate the structuring step of DRSA.

Alt.	g_1	g_2	D
a_1	3	2	Cl_0
a_2	5	1	Cl_0
a_3	5	1	Cl_0
a_4	3	2	Cl_1
a_5	4	3	Cl_1

2.1.2. Inference of Decision Rules

In the next step, knowledge is extracted from the already structured data. For this purpose, DRSA is coupled with decision rules, capturing the conditions capable of distinguishing the classes, hence relating performances on one or more criteria to a desired category. In the most common scenario that we implement in this work, the analysis involves inference of certain rules based on the lower approximations. This is due to their high reliability. In the case of non-monotonic criteria and binary classification, such rules take the following form: $g_{j_1}(a_i) \geq r_{j_1} \wedge \ldots \wedge g_{j_k}(a_i) \leq r_{j_k} \wedge \ldots \Rightarrow a_i \in Cl_r$, where $g_{j_1}, \ldots, g_{j_k}, \ldots \in G$ and $r_{j_1}, \ldots, r_{j_k}, \ldots$ correspond to the performances of alternatives contained in $\underline{P}(Cl_r)$.

To induce decision rules, we use DOMLEM (see Algorithm 1) [37]. It is applied separately to each decision class. The procedure is based on a greedy approach, iteratively constructing a conjunction of conditions to cover only alternatives from the given approximation. To select the best condition, firstly, its credibility, expressed as the fraction of correctly classified alternatives, is used. In case of a tie, generality, expressed as the total number of covered alternatives, is employed. Once the rule covering only alternatives from the lower approximation is created, a search for redundant conditions is performed. If, after removing some condition, the rule still covers only the desired alternatives, then such a condition is removed to ensure its minimality. The procedure of creating new rules is continued until all alternatives from the lower approximation are covered. Finally, if after removing some rule, the remaining ones still cover all alternatives from the lower approximation, then the rule is removed to ensure the minimality of the inferred set.

To demonstrate the operating steps of DOMLEM, we refer to a decision matrix given in Table 2. There are four alternatives in class Cl_0 ($a_1 - a_4$) and three alternatives in class Cl_1 ($a_5 - a_7$). For both classes, the lower and upper approximations are equal. Let us first infer certain rules for Cl_1. In the initial iteration, there are six potential conditions: $g_1 \geq 2$, $g_1 \geq 1$, $g_1 \leq 2$, $g_1 \leq 1$, $g_2 \geq 2$, and $g_2 \leq 2$. Among them, $g_2 \geq 2$ has the highest credibility of $\frac{3}{4}$ since it covers four alternatives, of which three belong to Cl_1. Therefore, it is added to the working condition part of the rule. Since it does not exclusively cover alternatives from the lower approximation of Cl_1, further conditions have to be included. There are five possible conditions left. Each is evaluated only on the set of alternatives covered by the constructed rule (in this case, $a_4 - a_7$). For example, condition $g_1 \geq 2$ covers two out of four alternatives (a_5 and a_7). In this iteration, conditions $g_1 \geq 1$ and $g_1 \geq 2$ share the highest credibility of one, but the former covers three alternatives while the latter only two. The conjunction of $g_2 \geq 2$ and $g_1 \geq 1$ covers only alternatives from $\underline{P}(Cl_1)$, so there is no need to add further conditions. Moreover, eliminating either condition will result in a rule covering alternatives from Cl_0. Thus, rule "if $g_2 \geq 2$ and $g_1 \geq 1$, then Cl_1" is minimal. It also covers all alternatives from $\underline{P}(Cl_1)$, so no more rules need to be inferred.

Algorithm 1 Induction of a set of certain rules using the DOMLEM algorithm

Input:
$\underline{P}$—a lower approximation from which the rules are to be induced
A^R—a reference set
Output:
R—a minimal set of certain decision rules
Procedures:
$alternatives(r)$—returns a set of alternatives covered by rule r
$eval(c)$—evaluates a set of conditions c using two metrics: credibility and generality

$leftToCover \leftarrow \underline{P}$
$R \leftarrow \emptyset$
while $|leftToCover| > 0$ **do**
  $Rule \leftarrow \emptyset$
  repeat
    $best \leftarrow None$
    for $g_j \in G$ **do**
      for $x \in \{g_j(a) \text{ for } a \in leftToCover \setminus alternatives(Rule)\}$ **do**
        ▷ Greedy search for the best condition
        for $d \in \{\leq, \geq\}$ **do**
          $condition \leftarrow \{g_j\ d\ x\}$
          ▷ Credibility and generality are compared lexicographically
          if $eval(condition \cup Rule) > eval(best \cup Rule)$ **then**
            $best \leftarrow condition$
          end if
        end for
      end for
    end for
    $Rule \leftarrow Rule \cup best$
  until $alternatives(Rule) \subseteq \underline{P}$
  for $condition \in Rule$ **do**
    if $alternatives(Rule \setminus condition) \subseteq \underline{P}$ **then**
      $Rule \leftarrow Rule \setminus condition$
      ▷ If a condition is redundant, it is removed
    end if
  end for
  $R \leftarrow R \cup \{Rule\}$
  $leftToCover \leftarrow leftToCover \setminus alternatives(Rule)$
end while
for $Rule \in R$ **do**
  if $\underline{P} = alternatives(R \setminus Rule)$ **then**
    $R \leftarrow R \setminus Rule$
    ▷ If a rule is redundant, it is removed
  end if
end for

Table 2. Example decision matrix used to illustrate the operating steps of DOMLEM.

Alt.	g_1	g_2	D
a_1	1	0	Cl_0
a_2	2	0	Cl_0
a_3	0	0	Cl_0
a_4	0	2	Cl_0
a_5	2	2	Cl_1
a_6	1	2	Cl_1
a_7	2	2	Cl_1

The same procedure is repeated for class Cl_0. Taking into account the unique performances on g_1 and g_2 of four alternatives contained in $\underline{P}(Cl_0)$, ten conditions are considered in the first iteration. Among them, $g_2 \leq 0$ is the best since it has the highest possible credibility of one and covers three alternatives. In comparison, $g_1 \leq 0$ also has the credibility of one but covers only two alternatives. Rule "if $g_2 \leq 0$, then Cl_0" covers three desired alternatives ($a_1 - a_3$), being minimal. To infer another rule covering a_4, we consider four possible conditions and favor $g_1 \leq 0$ with the credibility of one. Rule "if $g_1 \leq 0$, then Cl_0" covers two desired alternatives (a_3 and a_4), being minimal. Moreover, the two rules are needed to cover all alternatives in $\underline{P}(Cl_0)$, so they are both included in the returned set. Note that the rules can overlap. In this case, both rules cover alternative a_3. Therefore, the sum of the coverage coefficients of all rules can be greater than the size of the data set. However, the algorithm ensures that the inferred set of rules is minimal in the sense that eliminating any of them would leave at least one alternative uncovered.

2.1.3. Classification of Non-Reference Alternatives

The induced rules help classify non-reference alternatives from set A^{NR}. For each $a \in A^{NR}$, we identify all rules matching the performances of a. Then, three possible scenarios are possible in the case of binary classification [39]:

- No matching rule—an alternative is assigned to the majority class;
- One matching rule—an alternative is assigned to the class from the rule's decision part;
- A set of multiple matching rules $R = \{r_i, 1 \leq i \leq l\}$ with $|R| > 1$—an alternative is assigned to class Cl_i with the highest value of $score_R(Cl_i, a)$, according to the algorithm presented in [38]:

$$score_R(Cl_i, a) = \frac{|(Cond_{r_1} \cap Cl_i) \cup (Cond_{r_2} \cap Cl_i) \cup \cdots \cup (Cond_{r_l} \cap Cl_i)|^2}{|Cond_{r_1} \cup Cond_{r_2} \cup \cdots \cup Cond_{r_l}||Cl_i|},$$

where $Cond_{r_i}$ is a set of alternatives from A^R covered by rule r_i.

Section 3 provides an illustrative example of classification.

3. Results and Discussion

3.1. Overview of Decision Rules Obtained with DRSA

Each DRSA model was trained on 3515 compounds, which constitutes 80% of considered compounds. The rest was held out and used for model quality estimation. There were three independent resampling procedures of the non-drug set, each followed by the application of the DRSA rule inference algorithm. The drug molecule counterpart was kept the same. For example, the resulting $DRSA_1$ model comprises 402 decision rules (they are all provided in the online Supplementary Materials). Among them, 164 are responsible for drug case classification, whereas the remaining 238 support the non-drug cases. Although the number of rules is relatively high, they are simpler than the already-known alternatives. Most are based on three or fewer conditions while using at most two descriptors. Only around 15% of rules involve more than four conditions and less than 3% more than five conditions. Some are generic enough to support hundreds of cases, while others cover only a few training observations. Removing the weakest rules simplifies the DRSA model without significantly influencing the performance on a test set. For instance, excluding the rules covering at most five training examples reduces the total number of rules from 402 to 275, with only a slight effect of the sensitivity decrease by 1%.

The condition parts of the most supported rules for each class, along with their coverage, are presented in Table 3. Each row of the table represents a conjunction of conditions on the upper or lower limits of descriptor values that should be coupled with a

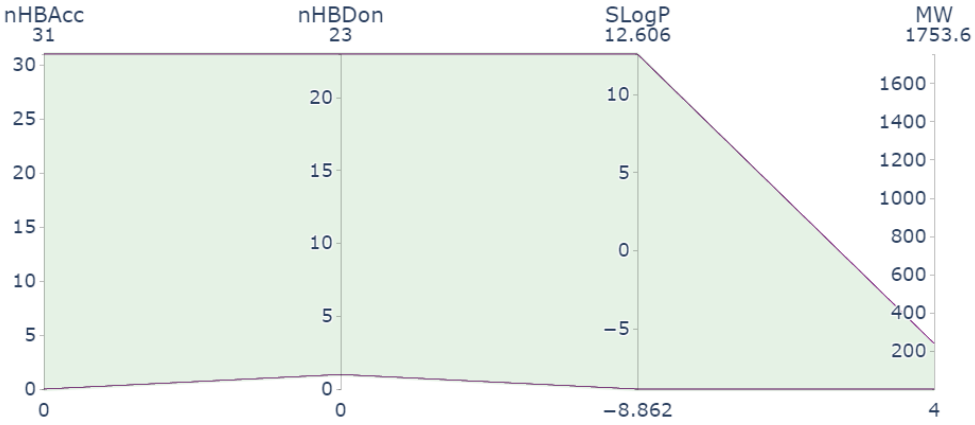
decision part in the form of either "then *drug*" or "then *non-drug*". In addition, the strongest drug rule (the first row in Table 3a)

$$\text{if MW} \leq 242.1 \text{ and nHBDon} \geq 1, \text{then } \textit{drug},$$

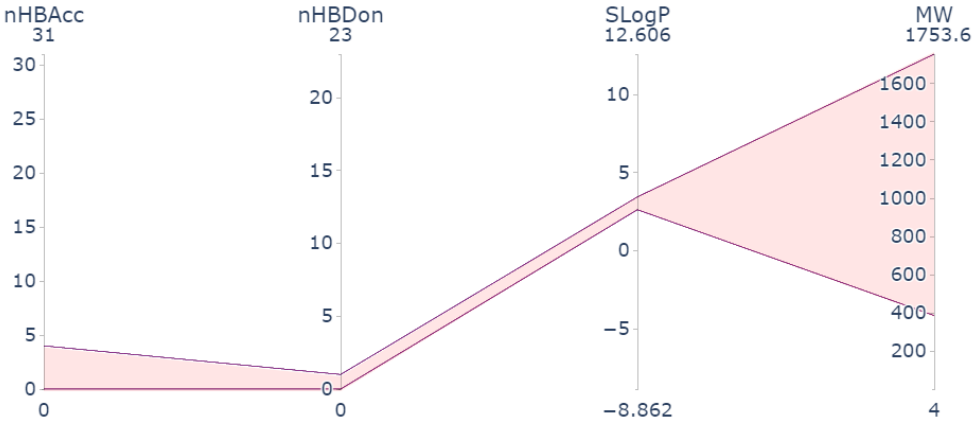
covering 302 drug cases is presented in Figure 1a, and the strongest non-drug rule (the first row in Table 3b)

$$\text{if MW} \geq 388.2 \text{ and } 2.6 \leq \text{SLogP} \leq 3.5 \text{ and nHBAcc} \leq 4 \text{ and nHBDon} \leq 1, \text{then } \textit{non-drug},$$

covering 143 non-drug cases is shown in Figure 1b.



(**a**) The strongest rule for a drug case.



(**b**) The strongest rule for a non-drug case.

Figure 1. Graphical representations of the strongest rules for both classes. The colored areas reflect the supported descriptor ranges.

Table 3. The strongest rules induced for both classes.

MW≥	MW≤	SLogP≥	SLogP≤	nHBDon≥	nHBDon≤	nHBAcc≥	nHBAcc≤	Coverage
(a) *Drug case rules* with a decision part "then *drug*"								
-	242.1	-	-	1	-	-	-	302
-	284.1	-	1.8	1	-	-	-	300
379.2	-	4.2	-	-	-	-	-	294
-	311.1	-	1.3	-	-	-	-	294
-	284.1	-	1.4	-	-	-	-	277
(b) *Non-drug case rules* with a decision part "then *non-drug*"								
388.2	-	2.6	3.5	-	1	-	4	143
398.2	-	2.6	3.7	-	1	-	4	137
395.1	429.1	2.6	3.5	-	1	-	5	132
397.1	434.1	-	3.5	-	1	-	4	119
337.3	435.1	-	3.5	-	-	4	5	118

The first *drug* rule imposes restriction exclusively on two descriptors: nHBAcc and nHBDon. It supports the cases with relatively low molecular weights and at least one HB donor. The second *drug* rule, of almost the same strength, additionally involves the octanol–water partition coefficient and enforces moderate lipophilicity to classify a molecule as a drug. On the other hand, the strongest *non-drug* rule involves all four descriptors. It assumes that the compounds of non-drug nature are relatively heavy (MW $\geq$ 388.2), of moderate lipophilicity lying in a relatively narrow range, and have at most 1 and 4 HB donors and acceptors, respectively. Some of these conditions are intuitively obvious. For instance, molecular weight apparently influences the classification process, which is clearly reflected in non-overlapping conditions involving this attribute. Similarly, restricted ranges of the number of HB donors and acceptors of the strongest *non-drug* rule can be related to severe limitations to establishing the HB interactions between these compounds and enzymes or receptors, which is often necessary to achieve biological activity. Interpreting some other conditions, e.g., those involving narrow ranges of the SlogP descriptor, is more challenging and requires training examples.

Independent of this, the predictions based on decision rules can easily be understood and interpreted. Their transparent nature supports the trust, possibly encouraging society to adopt the use of DRSA models. In particular, each prediction can be connected to the analysis of supporting rules. For example, let us consider the case of Xylometazoline, a widely applied drug that reduces the symptoms of nasal congestion. The positive prediction of the DRSA model is based on the following pair of rules:

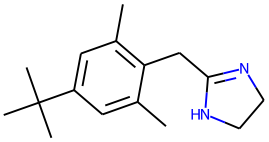
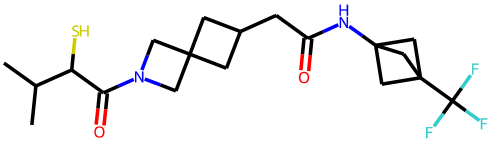
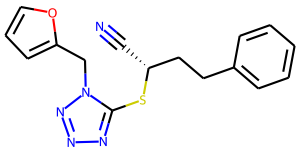
- if MW $\leq$ 266.2 and SLogP $\leq$ 3.7 and nHBAcc $\leq$ 2, then *drug*;
- if MW $\leq$ 255.1 and nHBAcc $\leq$ 2, then *drug*.

A non-drug example, properly classified by the DRSA model, is supported by the following three rules:

- if 398.2 $\leq$ MW $\leq$ 406.1 and SLogP $\leq$ 3.7 and nHBAcc $\leq$ 4 and nHBDon $\leq$ 3.0 , then *non-drug*;
- if 403.2 $\leq$ MW $\leq$ 461.1 and SLogP $\leq$ 3.7 and nHBAcc $\leq$ 3, then *non-drug*;
- if 379.2 $\leq$ MW $\leq$ 408.2 and 3 $\leq$ SLogP $\leq$ 3.3 and nHBAcc $\leq$ 3, then *non-drug*.

Table 4 shows the descriptors, structures, and predictions delivered by the DRSA, Lipinski, and Ghose models for both structures (upper and middle parts). In the case of the non-drug example, only DRSA is successful. The Lipinski and Ghose approaches are too liberal to exclude this compound from the set of potential drugs.

Table 4. Three exemplary compounds from the test set representing the drug and non-drug cases. The upper part reflects the drug molecule, whereas the middle and bottom sections represent the non-drug examples.

nHBAcc	nHBDon	SLogP	MW	P_{DRSA}	P_{Lip}	P_{Gho}
Example drug: Xylometazoline (CHEMBL312448)						
2	1	3.15	244.19	1	1	1
Upper						
Example non-drug: ZINCrB000003GBgp						
3	2	3.17	404.17	0	1	1
ine Middle						
Complex example non-drug: ZINCnz000000hm6C						
7	0	2.93	325.1	0	1	1
Bottom						

Nonetheless, the characteristics of a compound may also align with the rules supporting conflicting assignments. For illustrative purposes, let us consider a molecule presented at the bottom of Table 4 consistent with the following three rules:

- if MW $\leq$ 346.1 and nHBAcc $\geq$ 7 and nHBDon $=$ 0, then *drug*;
- if MW $\leq$ 331.1 and SLogP $\geq$ 1.3 and nHBAcc $\geq$ 7, then *drug*;
- if MW $\leq$ 399.1 and 2.9 $\leq$ SLogP $\leq$ 3.2 and nHBAcc $\geq$ 6, then *non-drug*.

The rules covering the alternative are used for computing the scores reflecting the support for assigning a compound to each class. The score is calculated by dividing the squared number of examples from the given class covered by these rules by the product of the total number of examples covered by these rules and the size of the given class in the reference set A^R [38]. For the considered compound, the two positive rules jointly cover 9 drugs out of 1767 in the training set. The non-drug rule covers 19 compounds out of 1748 non-drugs. Thus, the score for the class "drug" is equal to $\frac{9^2}{9*1767} = 0.0051$, and for the class "non-drug"—$\frac{19^2}{19*1748} = 0.011$. Finally, we recommend the class with the highest score. Therefore, the considered molecule is classified as a non-drug, and again, only the DRSA model predicts the non-drug nature of this compound properly.

3.2. Comparison of DRSA with Domain-Specific Methods

As demonstrated in Table 5, DRSA is superior with respect to the reference approaches proposed by Lipinski [1] and Ghose [3]. In both cases, DRSA reaches all classification quality metrics at the level of at least 80%. For example, the models run on the same attributes as the Ghose filter, which correctly indicated the "drug" class for 97% of all molecules they predicted as drugs, as captured by *precision*. They also correctly predicted the class for 91% of molecules overall, whether they were drugs or not, as reflected by *accuracy*.

Table 5. Classification quality measure comparison between the DRSA method, the reference approaches of Lipinski RO5 and Ghose filter, and selected ML classification methods. The columns with the "avg" subscript contain the average of three runs with different data samples. All results were obtained with a randomly chosen held-out data set containing 20% of all observations. The bold font indicates the best algorithm given a specific measure.

	Run on the same attributes as applied in the Lipinski RO5										
Measure	$DRSA_1$	$DRSA_2$	$DRSA_3$	$DRSA_{avg}$	Lip_{avg}	XGB_{avg}	NN_{avg}	SVM_{avg}	KNN_{avg}	AML_{avg}	ChP_{avg}
Sensitivity	0.80	0.80	0.80	0.80	0.78	**0.87**	0.85	0.80	0.82	**0.87**	0.91
Specificity	0.94	0.95	0.94	**0.95**	0.01	0.91	0.89	0.92	0.92	0.92	0.94
Precision	0.93	0.94	0.93	**0.93**	0.43	0.91	0.89	0.91	0.91	0.91	0.94
Accuracy	0.87	0.88	0.87	0.87	0.39	**0.89**	0.87	0.86	0.87	**0.89**	0.93
F_1 score	0.86	0.87	0.86	0.86	0.55	**0.89**	0.87	0.85	0.86	**0.89**	0.93
	Run on the same attributes as applied in the Ghose filter										
Measure	$DRSA_1$	$DRSA_2$	$DRSA_3$	$DRSA_{avg}$	$Ghose_{avg}$	XGB_{avg}	NN_{avg}	SVM_{avg}	KNN_{avg}	AML_{avg}	ChP_{avg}
Sensitivity	0.83	0.85	0.85	0.84	0.62	**0.89**	0.85	0.75	0.85	0.87	0.91
Specificity	0.96	0.98	0.98	**0.97**	0.07	0.96	0.91	0.96	**0.97**	0.95	0.94
Precision	0.95	0.97	0.97	**0.97**	0.39	0.96	0.91	0.95	**0.97**	0.94	0.94
Accuracy	0.90	0.91	0.92	0.91	0.34	**0.92**	0.88	0.85	0.91	0.91	0.93
F_1 score	0.89	0.91	0.91	0.91	0.49	**0.92**	0.88	0.84	0.90	0.90	0.93

On the contrary, the Lipinski RO5 and Ghose rules perform poorly on these real-world data. Both marked a vast majority of the non-drug molecules as active. In particular, the Lipinski rules misclassified over 99% non-drug examples, hence attaining extremely low scores on the specificity metric. In turn, the Ghose filter attained less favorable results on the remaining four metrics (e.g., its sensitivity was lower by 0.16 while its F_1 score was worse by 0.06). Such unfavorable performance of the classical techniques could be anticipated given the simplicity of individual rules defined by only four conditions.

The DRSA rules are more capacious and thus capable of extracting the patterns from provided data more thoroughly. The DRSA models perform well in specificity, while a higher false negative rate causes a slightly worse absolute sensitivity value (though still better than for the classical rules). This means that DRSA will tend to miss a small fraction of real drugs rather than classifying non-drug molecules as drugs. Such conditioning of the problem has certain benefits. Massive application of the DRSA model to a large number of molecules will less often bring false positive cases, which may reduce the costs of experimental verification of the molecule's activity. Interestingly, DRSA attained better results on all metrics when run on the attribute set applied by the Ghose procedure rather than Lipinski RO5. For example, the average accuracy increased from 0.87 to 0.91 while the F_1 score rose from 0.86 to 0.91. This shows further potential for improving the method's performance when supplied with richer, more adequate information.

The DRSA models can easily be applied to support the early-stage drug discovery process. As with any other statistical model, it can be trained on a selected set of compounds for which the experimental data exist. Subsequently, carefully created and—what is equally essential—thoroughly validated models can expose new, unseen compounds to deliver predictions and their intuitive interpretations. Due to relative numerical simplicity and clear potential for parallelization, these models can handle ultra-large, enumerated chemical databases containing synthesizable compounds, such as Enamine REAL, WuXi, or e-Molecules.

It is also worth mentioning that the presented methodology can naturally address new chemical modalities involving the compounds that are beyond Rule of Five (bRO5). In particular, DRSA can be applied to data associated with RNA therapeutics, protein degraders, cyclopeptides, and other classes of molecules [40–42]. The obtained rules, accounting for compounds from expanded chemical space, can be applied to, e.g., virtual screening of databases or used as criticizers of new molecules obtained in the frame of various generative chemistry approaches.

3.3. Comparison of DRSA with Selected Machine Learning Methods

We also confronted the predictive capabilities of DRSA with six selected popular ML methods. These included the following:

- eXtreme Gradient Boosting (XGB) [43]—a powerful ML algorithm based on decision trees, using gradient boosting to optimize performance; it focuses on speed and performance through regularization and efficient parallelization; we employed a variant with the number of trees between 10 and 250.
- Neural Network (NN) [44]—a classical multilayer perceptron with two hidden layers with the number of neurons between 32 and 512 each, *relu* activation, *adam* solver, and L2 regularization.
- Support Vector Machine (SVM) [45]—a supervised ML algorithm mapping input data into a higher-dimensional space using a kernel function (in our case, a radial variant), identifying the optimal hyperplane that maximizes the margin between different classes, and using this hyperplane to classify new data points by determining which side of the margin they fall on.
- K-nearest neighbors (KNN) [46]—a classification method calculating the distance (in our case, Euclidean) between the alternative to be classified and all reference alternatives and assigning it to the class most common among its nearest K neighbors (in our case, $3 \leq K \leq 19$).
- AutoML (AML) [47]—a procedure automatically adjusting the entire prediction pipeline, including a classification algorithm and a preprocessing phase using the genetic optimization procedure.
- ChemProp (ChP) [48]—a deep learning graph-convolutional Neural Network based on Message Passing Neural Network accompanied by a feature extraction module accepting a chemical molecule as an input. This algorithm uses the full molecule representation, not just simple descriptors like all remaining models used in this study. Therefore, it has an unfair advantage. For that reason, the metrics associated with the model should be considered as model performance at the prediction limit. They are not accounted for selecting the best score, and for clarity, they are separated by a vertical line.

The algorithms' hyperparameter values were determined using a Bayesian optimization procedure [49]. The use of all these approaches was preceded by the application of min–max scaling of performances.

The results are shown in Table 5. The proposed DRSA algorithm performs comparably to the state-of-the-art ML models. In both data sets, DRSA attained a very high score on specificity and precision while losing to XGB on sensitivity, accuracy, and F_1 score. Note that depending on the context, maximizing the score on a different metric might be desirable. Given the obtained results, DRSA seems to be more suitable when false positives are costly (specificity) or their cost is high (precision). In turn, XGB is more adequate when missing positive cases have high consequences (sensitivity), classes are well balanced and false positives and false negatives carry similar costs (accuracy), or there is class imbalance and you want a satisfying trade-off between precision and sensitivity (F_1).

We have also investigated the influence of feature space extension on the performance of the DRSA model as well as selected ML approaches. All considered methodologies were applied to data sets containing molecular descriptors related to both the Lipinski and Ghose filters, namely the number of HB donors, the number of HB acceptors, molecular weight, octanol–water partition coefficient (log P), molar refractivity, and the number of non-hydrogen atoms. The results are presented in Table 6. As expected, we observe an improvement in model quality in nearly all cases. Specifically, for the DRSA model, sensitivity and specificity increased by 0.06 and 0.01, respectively, compared to the DRSA results obtained using only RO5 descriptors. Similarly, for the ML models, we also note slightly better performance, except for the KNN model trained on Ghose descriptors, where sensitivity and specificity decreased slightly. Again, selecting the best model depends on which metric is the most important in a given scenario. There is not a single method that is

objectively the best on all metrics. Both DRSA and XGB attained the highest performance on different metrics, with the DRSA method having the additional advantage of easy interpretability. In conclusion, the DRSA models presented here demonstrate the expected behavior, with a vertical extension of the data leading to performance improvements.

Table 6. Classification quality measures of DRSA and selected ML models. In all cases, the quality measure reflects the mean value of three independent resampling procedures of the non-drug data set while the drug molecule counterpart is kept constant. The bold font indicates the best algorithm given a specific measure.

	DRSA	XGB	NN	SVM	KNN	AML	ChP
Sensitivity	0.86	**0.89**	0.86	0.81	0.84	0.83	0.91
Specificity	**0.96**	0.95	0.91	0.93	0.92	0.90	0.94
Precision	**0.96**	0.95	0.91	0.93	0.92	0.90	0.94
Accuracy	0.91	**0.93**	0.89	0.87	0.88	0.86	0.93
F_1 score	0.91	**0.93**	0.88	0.87	0.88	0.86	0.93

4. Summary

Our study proposed a new tool supporting the early stages of drug discovery. The DRSA method brings a vital instrument capable of creating well-performing, efficient, and easily interpretable models. Conceptually, they are of the same class as the Lipinski RO5 or Ghose filter. However, they are inferred systematically, assuring efficient knowledge extraction from data. Such rules can potentially improve the quality of the virtual search by increasing the quality of the filter or targeting specific molecular properties. In the application described here [50], the classical RO5 could be replaced with its DRSA analog, which could bring a better selection of the compounds for further processing. As an illustration, we learned the DRSA models from the original data set containing drug and non-drug molecules. The models' performance was estimated on the held-out data, proving superior to their classical counterparts and comparable with state-of-the-art machine learning models characterized by low interpretability capabilities. The straightforward nature of the inferred rules strongly supports understanding the predictions and encourages further applications in real-life scenarios. As a further study, we aim to apply the DRSA method to selected data sets containing the biological activity of various receptors and enzymes. We will also collect and explicitly include compounds related to new drug modalities, like peptides, macrocycles, or recently popular proteolysis-targeting chimeras (PROTACs).

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/app14219966/s1: A set of all drug and non-drug decision rules induced using the Dominance-based Rough Set Approach.

Author Contributions: Conceptualization, G.M., A.M., M.K. and R.A.B.; Methodology, G.M., A.M., M.K. and R.A.B.; Software, G.M., A.M. and R.A.B.; Validation, G.M. and R.A.B.; Formal analysis, G.M., M.K. and R.A.B.; Investigation, G.M., A.M. and R.A.B.; Resources, M.K.; Data curation, R.A.B.; Writing—original draft, G.M., A.M., M.K. and R.A.B.; Visualization, G.M. and R.A.B.; Supervision, M.K. and R.A.B.; Project administration, M.K. and R.A.B.; Funding acquisition, M.K. All authors have read and agreed to the published version of the manuscript.

Funding: G. Miebs and A. Mielniczuk were supported by the Polish Ministry of Science and Higher Education, grant no. 0311/SBAD/0741. M. Kadziński acknowledges support from the Polish National Science Center, grant no. DEC-2019/34/E/HS4/00045. R. Bachorz acknowledges support from the Polish National Science Center, grant no. 2019/33/B/NZ7/00795.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found at https://www.ebi.ac.uk/chembl/ (accessed on 28 October 2024) and https://cartblanche.docking.org/ (accessed on 28 October 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3–25. [CrossRef]
2. Lipinski, C.A. Lead- and drug-like compounds: The rule-of-five revolution. *Drug Discov. Today Technol.* **2004**, *1*, 337–341. [CrossRef] [PubMed]
3. Ghose, A.K.; Viswanadhan, V.N.; Wendoloski, J.J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1999**, *1*, 55–68. [CrossRef] [PubMed]
4. Zheng, S.; Luo, X.; Chen, G.; Zhu, W.; Shen, J.; Chen, K.; Jiang, H. A New Rapid and Effective Chemistry Space Filter in Recognizing a Druglike Database. *J. Chem. Inf. Model.* **2005**, *45*, 856–862. [CrossRef]
5. Chen, G.; Zheng, S.; Luo, X.; Shen, J.; Zhu, W.; Liu, H.; Gui, C.; Zhang, J.; Zheng, M.; Puah, C.M.; et al. Focused Combinatorial Library Design Based on Structural Diversity, Druglikeness and Binding Affinity Score. *J. Comb. Chem.* **2005**, *7*, 398–406. [CrossRef]
6. Rishton, G.M. Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discov. Today* **2003**, *8*, 86–96. [CrossRef]
7. Veber, D.F.; Johnson, S.R.; Cheng, H.Y.; Smith, B.R.; Ward, K.W.; Kopple, K.D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615–2623. [CrossRef]
8. Walters, W.; Murcko, M.A. Prediction of 'drug-likeness'. *Adv. Drug Deliv. Rev.* **2002**, *54*, 255–271. [CrossRef]
9. Hann, M.M.; Leach, A.R.; Harper, G. Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856–864. [CrossRef]
10. Oprea, T.I.; Gottfries, J.; Sherbukhin, V.; Svensson, P.; Kühler, T.C. Chemical information management in drug discovery: Optimizing the computational and combinatorial chemistry interfaces. *J. Mol. Graph. Model.* **2000**, *18*, 541. [CrossRef]
11. Oprea, T.I. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251–264. :1008130001697 [CrossRef] [PubMed]
12. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'Rule of Three' for fragment-based lead discovery? *Drug Discov. Today* **2003**, *8*, 876–877. [CrossRef]
13. Monge, A.; Arrault, A.; Marot, C.; Morin-Allory, L. Managing, profiling and analyzing a library of 2.6 million compounds gathered from 32 chemical providers. *Mol. Divers.* **2006**, *10*, 389–403. [CrossRef] [PubMed]
14. Benet, L.Z.; Hosey, C.M.; Ursu, O.; Oprea, T.I. BDDCS, the Rule of 5 and drugability. *Adv. Drug Deliv. Rev.* **2016**, *101*, 89–98. [CrossRef] [PubMed]
15. Bickerton, G.R.; Paolini, G.V.; Besnard, J.; Muresan, S.; Hopkins, A.L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **2012**, *4*, 90–98. [CrossRef]
16. Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325–3329. [CrossRef]
17. Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1882–1889. [CrossRef]
18. Takaoka, Y.; Endo, Y.; Yamanobe, S.; Kakinuma, H.; Okubo, T.; Shimazaki, Y.; Ota, T.; Sumiya, S.; Yoshikawa, K. Development of a Method for Evaluating Drug-Likeness and Ease of Synthesis Using a Data Set in Which Compounds Are Assigned Scores Based on Chemists' Intuition. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1269–1275. [CrossRef]
19. Ajay; Walters, W.P.; Murcko, M.A. Can We Learn To Distinguish between "Drug-like" and "Nondrug-like" Molecules? *J. Med. Chem.* **1998**, *41*, 3314–3324. [CrossRef]
20. Hooshmand, S.A.; Jamalkandi, S.A.; Alavi, S.M.; Masoudi-Nejad, A. Distinguishing drug/non-drug-like small molecules in drug discovery using deep belief network. *Mol. Divers.* **2021**, *25*, 827–838. [CrossRef]
21. Wagener, M.; Van Geerestein, V.J. Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 280–292. [CrossRef] [PubMed]
22. Schneider, N.; Jäckels, C.; Andres, C.; Hutter, M.C. Gradual in Silico Filtering for Druglike Substances. *J. Chem. Inf. Model.* **2008**, *48*, 613–628. [CrossRef]
23. Zernov, V.V.; Balakin, K.V.; Ivaschenko, A.A.; Savchuk, N.P.; Pletnev, I.V. Drug Discovery Using Support Vector Machines. The Case Studies of Drug-likeness, Agrochemical-likeness, and Enzyme Inhibition Predictions. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2048–2056. [CrossRef]
24. Gaulton, A.; Bellis, L.J.; Bento, A.P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107. [CrossRef]
25. Tingle, B.I.; Tang, K.G.; Castanon, M.; Gutierrez, J.J.; Khurelbaatar, M.; Dandarchuluun, C.; Moroz, Y.S.; Irwin, J.J. ZINC-22-A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery. *J. Chem. Inf. Model.* **2023**, *63*, 1166–1176. [CrossRef]
26. Pawlak, Z. Rough sets. *Int. J. Comput. Inf. Sci.* **1982**, *11*, 341–356. [CrossRef]
27. Greco, S.; Matarazzo, B.; Slowinski, R. Rough sets theory for multicriteria decision analysis. *Eur. J. Oper. Res.* **2001**, *129*, 1–47. [CrossRef]

28. Couto, A.B.G.D.; Gomes, L.F.A.M. Sovereign Rating Analysis through the Dominance-Based Rough Set Approach. *Found. Comput. Decis. Sci.* **2020**, *45*, 3–16. [CrossRef]
29. Oppio, A.; Dell'Ovo, M.; Torrieri, F.; Miebs, G.; Kadziński, M. Understanding the drivers of Urban Development Agreements with the rough set approach and robust decision rules. *Land Use Policy* **2020**, *96*, 104678. [CrossRef]
30. Boggia, A.; Rocchi, L.; Paolotti, L.; Musotti, F.; Greco, S. Assessing Rural Sustainable Development potentialities using a Dominance-based Rough Set Approach. *J. Environ. Manag.* **2014**, *144*, 160–167. [CrossRef]
31. Cinelli, M.; Spada, M.; Kadziński, M.; Miebs, G.; Burgherr, P. Advancing Hazard Assessment of Energy Accidents in the Natural Gas Sector with Rough Set Theory and Decision Rules. *Energies* **2019**, *12*, 4178. [CrossRef]
32. Barbati, M.; Greco, S.; Kadziński, M.; Słowiński, R. Optimization of multiple satisfaction levels in portfolio decision analysis. *Omega* **2018**, *78*, 192–204. [CrossRef]
33. Maaroof, N.; Moreno, A.; Valls, A.; Jabreel, M.; Szeląg, M. A Comparative Study of Two Rule-Based Explanation Methods for Diabetic Retinopathy Risk Assessment. *Appl. Sci.* **2022**, *12*, 3358. [CrossRef]
34. Błaszczyński, J.; Greco, S.; SŁowiński, R. Inductive discovery of laws using monotonic rules. *Eng. Appl. Artif. Intell.* **2012**, *25*, 284–294. [CrossRef]
35. Moriwaki, H.; Tian, Y.S.; Kawashita, N.; Takagi, T. Mordred: A molecular descriptor calculator. *J. Cheminform.* **2018**, *10*, 4. [CrossRef]
36. Landrum, G. RDKit: Open-Source Cheminformatics Software. 2022. Available online: http://www.rdkit.org (accessed on 28 October 2024).
37. Greco, S.; Matarazzo, B.; Slowinski, R.; Stefanowski, J. An Algorithm for Induction of Decision Rules Consistent with the Dominance Principle. In *Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2001; pp. 304–313. [CrossRef]
38. Błaszczyński, J.; Greco, S.; Słowiński, R. Multi-criteria classification—A new scheme for application of dominance-based decision rules. *Eur. J. Oper. Res.* **2007**, *181*, 1030–1044. [CrossRef]
39. Kadziński, M.; Greco, S.; Słowiński, R. Robust Ordinal Regression for Dominance-based Rough Set Approach to multiple criteria sorting. *Inf. Sci.* **2014**, *283*, 211–228. [CrossRef]
40. Blanco, M.J.; Gardinier, K.M. New Chemical Modalities and Strategic Thinking in Early Drug Discovery. *ACS Med. Chem. Lett.* **2020**, *11*, 228–231. [CrossRef]
41. Doak, B.; Over, B.; Giordanetto, F.; Kihlberg, J. Oral Druggable Space beyond the Rule of 5: Insights from Drugs and Clinical Candidates. *Chem. Biol.* **2014**, *21*, 1115–1142. [CrossRef]
42. Doak, B.C.; Zheng, J.; Dobritzsch, D.; Kihlberg, J. How Beyond Rule of 5 Drugs and Clinical Candidates Bind to Their Targets. *J. Med. Chem.* **2016**, *59*, 2312–2327. [CrossRef]
43. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, New York, NY, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
44. McCulloch, W.S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **1943**, *5*, 115–133. [CrossRef]
45. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
46. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
47. Le, T.T.; Fu, W.; Moore, J.H. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics* **2020**, *36*, 250–256. [CrossRef]
48. Heid, E.; Greenman, K.P.; Chung, Y.; Li, S.C.; Graff, D.E.; Vermeire, F.H.; Wu, H.; Green, W.H.; McGill, C.J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **2023**, *64*, 9–17. [CrossRef]
49. Močkus, J. On Bayesian methods for seeking the extremum. In Proceedings of the Optimization Techniques IFIP Technical Conference Novosibirsk, Novosibirsk, Russia, 1–7 July 1974; Springer: Berlin/Heidelberg, Germany, 1975; pp. 400–404. [CrossRef]
50. Zhao, H.; Liu, J.; He, L.; Zhang, L.; Yu, R.; Kang, C. Virtual screening and molecular dynamics simulation for identification of natural antiviral agents targeting SARS-CoV-2 NSP10. *Biochem. Biophys. Res. Commun.* **2022**, *626*, 114–120. [CrossRef]