RegioSQM20: Improved prediction of the regioselectivity of electrophilic aromatic substitutions

Nicolai Ree,[†] Andreas H. Göller,[*,‡] and Jan H. Jensen[*,†]

†Department of Chemistry, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen Ø, Denmark

‡Bayer AG, Pharmaceuticals, R&D, Computational Molecular Design, 42096 Wuppertal, Germany

E-mail: andreas.goeller@bayer.com; jhjensen@chem.ku.dk

Abstract

We present RegioSQM20, a new version of RegioSQM (*Chem. Sci.* 2018, 9, 660), which predicts the regioselectivities of electrophilic aromatic substitution (EAS) reactions from the calculation of proton affinities. The following improvements have been made: The open source semiempirical tight binding program `xtb` is used instead of the closed source `MOPAC` program. Any low energy tautomeric forms of the input molecule are identified and regioselectivity predictions are made for each form. Finally, RegioSQM20 offers a qualitative prediction of the reactivity of each tautomer (low, medium, or high) based on the reaction center with the highest proton affinity. The inclusion of tautomers increases the success rate from 90.7% to 92.7%. RegioSQM20 is compared to two machine learning based models: one developed by Struble et al. (*React. Chem. Eng.* 2020, 5, 896) specifically for regioselectivity predictions of EAS reactions (WLN) and a more generally applicable reactivity predictor (IBM RXN) developed by Schwaller et al. (*ACS Cent. Sci.* 2019, 5, 1572). RegioSQM20 and WLN offers roughly the same success rates for the entire data sets (without considering tautomers), while WLN is many orders of magnitude faster. The accuracy of the more general IBM RXN approach is somewhat lower: 76.3%-85.0%, depending on the data set. The code is freely available under the MIT open source license and will be made available as a webservice (regiosqm.org) in the near future.

1

Introduction

Halogenated derivatives of heteroaromatics and benzene derivatives are often applied as substrates in carbon–carbon and carbon–heteroatom cross-coupling reactions[1,2] and are typically prepared by electrophilic aromatic substitution (EAS). However, it is often not a priori obvious at which position(s) halogenation will occur for compounds in the late stages of the synthesis that contain multiple (hetero)aromatic rings or in compounds that contain both heteroarene and benzene rings. Consequently, organic chemists tend to install the halogens early in the synthesis, thereby effectively eliminating a large number of otherwise promising synthetic routes. This is not only a problem for the synthetic chemists but also developers of software that automatically construct retrosynthetic routes.[3–9]

Several predictive tools have been developed to address this problem based on heuristics,[10] quantum chemical calculations (QM),[11] machine learning (ML),[12,13] or a combination of QM and ML.[14,15] One of these methods is the RegioSQM method developed by Kromann et al.[11] (referred to hereafter as RegioSQM18). RegioSQM18 uses the semiempirical PM3 method[16] and the COSMO continuum solvation model[17] implemented in the `MOPAC` program. `MOPAC` is a closed-source software package that is free to academics but not to industry, so we decided to investigate open source alternatives for further development of RegioSQM. In this paper, we show that the open source semiempirical software package `xtb` can be used in place of `MOPAC` without impacting the accuracy of the predictions. We go on to show that the accuracy can be increased by considering different tautomeric forms of the molecule and offer a qualitative prediction of the reactivity of each tautomer. Finally, we compare the accuracy of the new version (RegioSQM20) to two ML-based models for regioselectivity predictions.
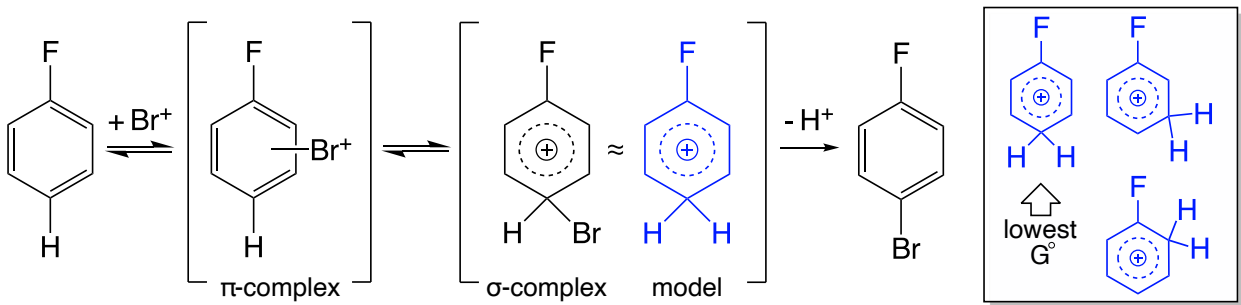

Figure 1: The mechanism of an electrophilic aromatic substitution (EAS) reaction using fluorobenzene as an example. RegioSQM approximates the Br σ-complex as protonation (shown as blue structures) and determines regioselectivity by finding the protonated isomer with lowest free energy.
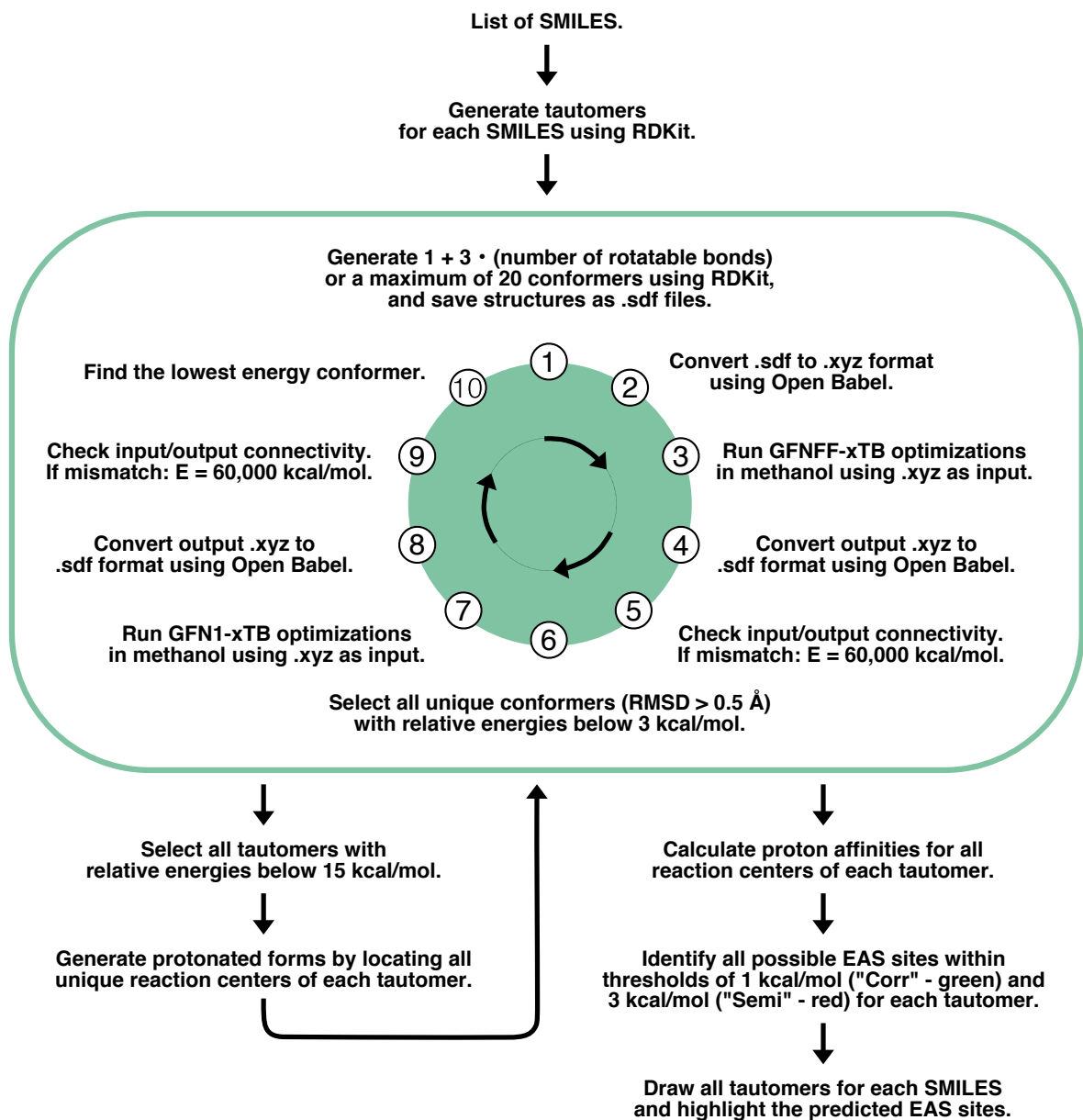
Figure 2: A flowchart describing the procedure of RegioSQM20.

Computational Methodology

Fig. 1 illustrates the EAS mechanism using the bromination of fluorobenzene as an example. The mechanism is relatively simple and involves the addition of an electrophile to the aromatic ring to form a σ-complex (also called a Wheland intermediate or an arenium ion), which usually determines the regioselectivity of EAS reactions with Br and Cl. Hence, free energy calculations of different protonated regioisomers, corresponding to different σ-complexes and thereby different reaction pathways, can be used to predict the regioselectivity.

The procedure of RegioSQM20 is fully automated with the only user input being a SMILES (simplified molecular input line entry system) representation of a given molecule (see Fig. 2). RegioSQM20 will then generate tautomers using TautomerEnumerator in RDKit 2020.03.1,[18] and all tautomers will go into a conformational search algorithm. In this algorithm, $\min(1 + 3 \cdot n_{\mathrm{rot}}, 20)$ conformers are generated for each tautomer, where n_{rot} is the number of rotatable bonds. The conformers are then optimized in methanol (MeOH, dielectric = 33.6) using the fast force-field version of xTB called GFNFF-xTB and the generalized Born (GB) model with solvent accessible surface area (SASA) termed GBSA.[19] Compared to RegioSQM18, this new implementation generates conformers using EmbedMultipleConfs from RDKit with ETversion=2 instead of ETversion=1. After this procedure, all conformers with relative total energies below 3 kcal/mol are clustered with the Butina algorithm in RDKit to find unique conformers using the pairwise heavy-atom position root mean square deviation (RMSD) with a threshold of 0.5 Å. The cluster centroids are then re-optimized in MeOH using GFN1-xTB and the GBSA solvation model in order to find the lowest energy conformers.[20] After both optimizations, the input and output structures are compared by converting the Cartesian coordinate file (.xyz) into a structure-data file (.sdf) using Open Babel 2.4.1.[21] If the atom connectivity is different, due to e.g. an intramolecular proton transfer reaction or a broken/created bond, the energy of the molecule is set to 60,000 kcal/mol. In case the force-field calculation fails, the initial RDKit structure will be used as the input structure for the GFN1-xTB calculation. Hereafter, RegioSQM20 selects all tautomers with relative total energies below 15 kcal/mol and locates all unique reaction centers to generate single protonated forms of the tautomers. These protonated molecules are then sent into the conformational search algorithm to find their lowest energy conformer. Subsequently, the proton affinities are calculated as the energy difference between the unprotonated and protonated forms. Note that we neglect the energy of the proton in solution, since the qualitative reactivity categorization is based on comparing the proton affinities to cutoff values where this term is also neglected. The predicted EAS sites are then identified as the reaction centers with proton affinities within 1 kcal/mol ("Corr" - green) and 3 kcal/mol ("Semi" - red) of the highest proton affinity. Finally, the tautomers are drawn and their respective EAS sites are highlighted as seen in Fig. 3. All tautomers are labelled with the relative energies of the unprotonated forms and the predicted reactivity based on the highest proton affinity of each tautomer.

The source code is freely available on GitHub (https://github.com/jensengroup/RegioSQM20) and will be made available as a web service at molcalc.org in the near future.
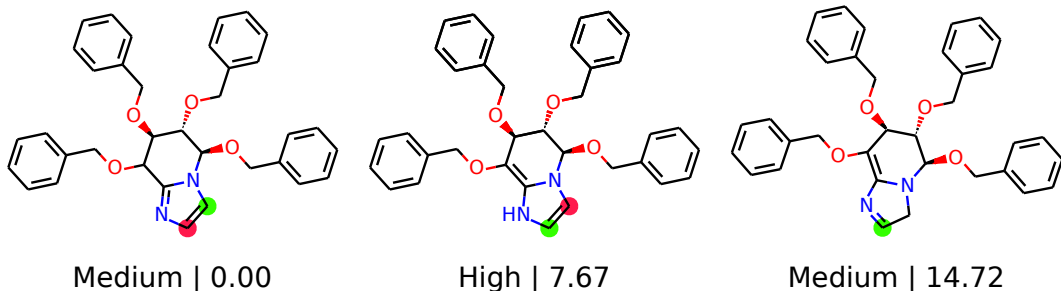
Figure 3: An example of the output of RegioSQM20. Tautomers are depicted with relative energies (only including those below 15 kcal/mol) along with an estimate of their reactivity based on the highest proton affinity. The highlighted atoms represent the predicted EAS sites within 1 kcal/mol (green circles) and the EAS sites within 3 kcal/mol (red circles).

Table 1: Comparing RegioSQM implementations. The last entry corresponds to RegioSQM20. For the generation of the conformers, a random seed of 90 was used. Furthermore, the RegioSQM18 and GFN1-xTB/methanol entries used ETversion=1 and otherwise ETversion=2.

Methodology	Corr/Semi/Fail	Median time (s)^a	Mean time (s)^a	Total time (h)^{a,b}
RegioSQM18	488/30/17	42	127	10
GFN1-xTB/methanol	486/27/22	60	230	17
FF Optimization	485/29/21	33	110	8
Tautomers (canonical RDKit)	477/28/30	33	110	8
Tautomers (lowest energy)	483/27/25	39	163	12
Tautomers (15 kcal/mol)	496/21/18	49	223	17

^{a}4 cores/molecule (Intel(R) Xeon(R) CPU X5550 @ 2.67GHz). ^{b}Two molecules running in parallel.
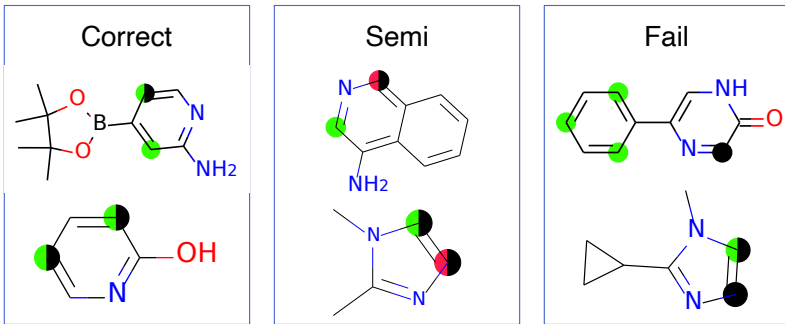
Figure 4: Definition of correct, semi-correct, and incorrect predictions of regioselectivity. RegioSQM correlates proton affinity with EAS reactivity. The green and red dots indicate the atoms with proton affinities within 1 and 3 kcal/mol of the highest value, respectively. The black dots indicate the experimentally observed sites for EAS. All experimentally observed sites must be predicted by green dots to be counted as correct.

Results and Discussion

Switching from PM3 to GFN-xTB

RegioSQM18 was developed and tested using 535 EAS reactions collected from the literature[11] and we use the same data set to guide the development of RegioSQM20. RegioSQM18 predicts the correct regioselectivity for 488 of the 535 reactions, while 30 and 17 are predicted semi-correctly and incorrectly, respectively (Table 1). A correct prediction is one where all experimentally observed sites have predicted proton affinities within 1 kcal/mol of the highest proton affinity (marked as green in Fig. 4). The definition of a semi-correct prediction is the same as a correct prediction except that the 1 kcal/mol cutoff is changed to 3 kcal/mol (marked as red in Fig. 4). Finally, an incorrect (or failed) prediction is one where at least one experimentally observed site is not predicted correctly. We repeated the calculations with GFN1-xTB and GFN2-xTB in combinations with a variety of solvents and found that GFN1-xTB and methanol gave the most accurate results with 486, 27, and 22 correct, semi-correct, and failed predictions (Table 1 and Fig. S1). This result is very similar to those obtained with PM3 and shows that an open source method can be used instead of PM3.

Use of the GFN1-xTB/GBSA method increases the median CPU time requirement from 43 to 60 seconds and the total time for all 535 increases from 10 to 17 hours. In order to mitigate the increased computational cost, we tested the use of the GFNFF-xTB force field to pre-screen the conformers. If only conformers with GFNFF-xTB/methanol energies within 3 kcal/mol of the lowest energy conformer are re-optimized using GFN1-xTB/methanol then the number of correct, semi-correct, and failed predictions (485, 29, and 21) are essentially unchanged (Table 1) while decreasing the median CPU time per molecule to 33 seconds and total time to 8 hours.

Tautomers

The experimental regioselectivity data we use is collected from the literature and the tautomeric forms of the molecules in that data set are those displayed in the respective publications. A few exploratory calculations revealed that the tautomeric form of the molecule can have an impact on the predicted regioselectivity, so we decided to address this issue in a more systematic fashion. If we instead use the tautomer with the lowest GFN1-xTB/methanol energy the number of correct predictions is 483 (Table 1), which is slightly lower than that obtained without considering tautomers. The most probable explanation is that GFN1-xTB/methanol is not sufficiently accurate to identify the most stable tautomer and/or that this property is more sensitive to the choice of solvent. We therefore investigated the effect of including all tautomers within a certain cutoff of the lowest energy form and the results are shown Fig. S2. The figure shows that a relatively large cutoff (>8 kcal/mol) is needed in order for the inclusion of tautomers to have a significant effect on the accuracy.

A correct prediction is one where all the experimentally observed sites are predicted (with a green dot) by at least one of the tautomers, so one worry is that the high cutoff simply leads

to a large number of tautomers each with a different site being predicted as most reactive. However, an analysis of the data (Fig. S3) shows that 91% of the molecules in our data set have only have one or two tautomers, even for a cutoff of 15 kcal/mol. This statistic is reflected in the 11 molecules for which the predictions improve using a cutoff of 15 kcal/mol (Fig. S4). Of these 11 molecules only two have more than two tautomers and in all cases the increase in accuracy is a result of only one new position being predicted as most reactive (Fig. S4). Thus, RegioSQM20 uses a 15 kcal/mol cutoff for tautomers, which increases the number of correct predictions to 496 molecules (Table 1) and the median and total CPU time to 49 seconds and 17 hours, respectively.

The CPU time can be significantly reduced by using GFN2-xTB with only a relatively minor decrease in accuracy. For example, GFN2-xTB/water has a median CPU time of only 29 s and total time of 10 hours, while the number of correct and incorrect predictions are 493 and 24, respectively. Thus, this method can be selected if computational efficiency is a greater concern than accuracy.

Table 2: Comparison of RegioSQM to two ML-based models. Tautomers are not considered so "RegioSQM20*" (note the "*") corresponds to the "FF Optimization" entry in Table 1. "One reactive atom" refers to the subset of the full data set with only one unique experimentally observed reaction site and where RegioSQM20* only predicts one unique (green) reactive site.

Methodology	Full data set Corr/Semi/Fail	One reactive atom Corr/Fail
RegioSQM20*	90.7%/5.4%/3.9%	92.0%/8.0%
WLN	89.1%/4.9%/6.0%	96.0%/4.0%
IBM RXN	76.3%/23.7%	85.0%/15.0%

Comparison to machine learning models

Several machine learning (ML) based models predict regioselectivity of EAS reactions and have been compared to RegioSQM18[7,12], such as the Weisfeiler-Lehman neural network (WLN) based method by Struble et al.[12] and the molecular transformer (MolTrans) by Schwaller et al.[7].

Struble et al. compared the top one, two, and three predictions of the WLN-based method, which predicts a reaction probability for each atom, to those of RegioSQM18 and found success rates of 85.0-95.7% for WLN, compared to 79.7-93.3% for RegioSQM18. However, this approach does not consider the differences in the reaction probabilities (in the case of WLN) nor relative energies of the protonated isomers (in the case of RegioSQM). For example, a top-2 prediction would be correct even if the second reaction probability is extremely low, like for instance 98% and 6% for the first and second position, and similarly

for protonated isomers with relative energies of 0 and 15 kcal/mol for RegioSQM. Conversely, a top-1 prediction would be considered incorrect even if the top two reaction probabilities are 99% and 98%, and the reaction is observed to occur at the site with a 98% reaction probability. Or, in the case of RegioSQM, if the reaction is observed to occur at a site where the corresponding isomer is 0.2 kcal/mol higher than the isomer with the lowest energy. Instead, we therefore define a correct WLN prediction if the observed reaction site has a reactivity score that is within 15% of the highest score, and a semi-correct otherwise as long as the prediction probability is >5% (the cutoff the authors used to classify non-reactive atoms). With these definitions, and without considering tautomers, the number of correct and incorrect WLN-based predictions are 477 and 32, respectively, compared to 485 and 21 for RegioSQM20* ("FF optimization" entry in Table 1, i.e. without considering tautomers). This corresponds to a success rate of 89.1% and 90.7% for WLN and RegioSQM20*, respectively, while the corresponding failure rates are 6.0% and 3.9% (Table 2). For comparison the success and failure rates for RegioSQM20 with tautomers are 92.7% and 3.4%, respectively.

Since the chosen cutoffs are somewhat arbitrary, we also investigate the subset of 426 molecules with only one unique experimentally observed reaction site and where RegioSQM20* only predicts one unique (green) reactive site. Here, RegioSQM20* is compared to the top-1 WLN prediction and the success rates for WLN and RegioSQM20* are 96.0% and 92.0%, respectively. A similar comparison to RegioSQM18 by Struble et al. for a different set of molecules (not used to train the WLN method) yielded 87.9% and 86.7%, respectively. The relatively large difference in success rates observed for WLN for these two sets of molecules could indicate that many of the molecules used in the current data set are included in the WLN-training set. Overall, the success rates of RegioSQM20 and the WLN-based method are thus comparable, while the latter is orders of magnitude faster.

The development of the techniques underlying MolTrans has been continued in the IBM RXN for Chemistry (IBM RXN) package,[22] so this is the package we compare to RegioSQM20. We use N-bromosuccinimide (NBS) as the reagent since this is the source of Br for most of the reactions in our data set and, as before, we use the tautomeric form found in the data set. With IBM RXN we only have access to the most likely prediction and with this limitation the success rate is 76.3% using the entire data set. This is somewhat lower than the top-1 success rate of 83% reported for MolTrans by Schwaller et al. for a different data set. It is not clear whether the difference is due to differences in the data set or differences between MolTrans and IBM RXN. However, the corresponding success rate for the 426 molecules with only one unique experimentally observed reaction site and where RegioSQM20* only predicts one unique (green) reactive site is 85.0%, which is closer to the value reported by Schwaller et al. and somewhat lower than the 96.0% and 92.0% observed for WLN and RegioSQM20*.

Prediction of reactivity

The original inspiration for using proton affinities to predict regioselectivity came from the observation by Streitwieser and others that the rates of many EAS reactions correlate well

with the proton affinity of the reacting carbon.[23] While RegioSQM18 only predicts the relative proton affinities (i.e. the relative energies of the protonated isomers), the proton affinity (i.e. the energy difference between the unprotonated and protonated forms) can be calculated at no additional cost since the energy of the unprotonated form of the molecule is computed to identify low energy tautomers. The proton affinities are more difficult to obtain accurately than relative proton affinities, so we only expect these values to give a qualitative indication of reactivity. Fig. 5a shows the highest proton affinity of the most stable tautomer computed for a series of substituted benzene analogs familiar to all organic chemists together with a qualitative ranking of their reactivities. There is a clear separation in the proton affinities of the most (80-91 kcal/mol) and least (51-67 kcal/mol) reactive molecules. An example from each of these classes along with the proton affinities of all the unique reaction centers can be seen in Fig. 5b.

Having established a qualitative correlation between reactivity and proton affinity computed by GFN1-xTB/methanol, we computed the highest proton affinity of the most stable tautomer for the 535 molecules in our data set. The results (5c) show that 92% of the molecules have proton affinities in the range 70-100 kcal/mol - a range similar to that found for most of the reactive (ortho-para directing) molecules shown in 5a. Molecules with higher and lower proton affinities are thus deemed unusually high and low reactivity, respectively. RegioSQM20 therefore uses these cutoffs to classify a molecule as having low, medium, or high reactivity and displays this information in the output (Fig. 5d).
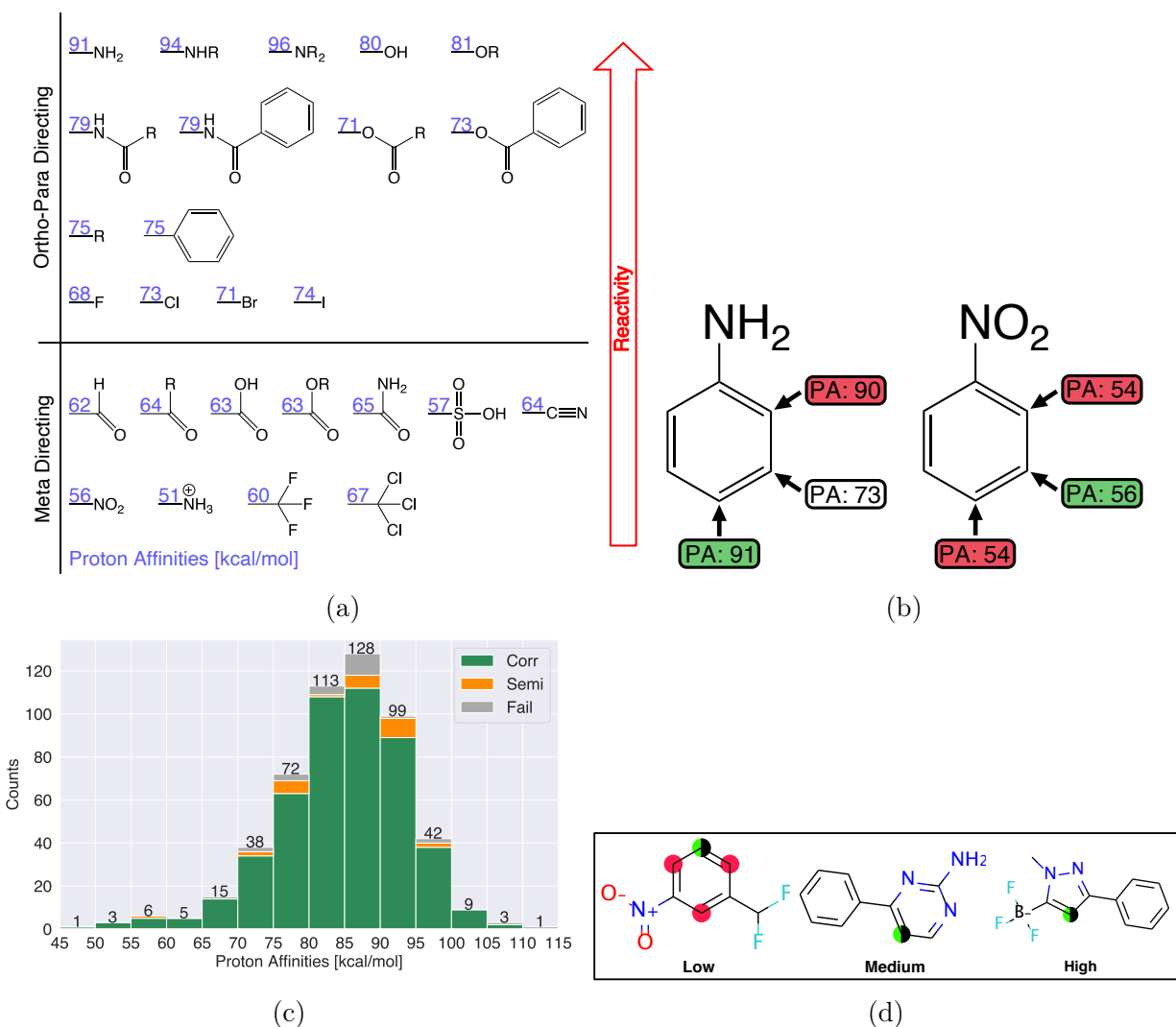
Figure 5: (a) The highest proton affinity of the most stable tautomer computed for a series of substituted benzene analogs familiar to all organic chemists together with a qualitative ranking of their reactivities. The values are given in units of kcal/mol. (b) Two examples from (a) showing the proton affinities of the unique reaction centers. (c) Computed proton affinities for the 535 molecules in our data set as the energy difference between the unprotonated and most stable protonated form using the most stable tautomer. (d) Examples of molecules with predicted low, medium, and high reactivity based on the highest proton affinities.

Conclusions and outlook

We present RegioSQM20, a new version of RegioSQM[11], which predicts the regioselectivities of electrophilic aromatic substitution (EAS) reactions by finding the atomic center with the highest proton affinity. The following improvements have been made: The open source semiempirical tight binding program `xtb` is used instead of the closed source `MOPAC` program; specifically GFN1-xTB/methanol is used instead of PM3/chloroform. Any low energy tau-

tomeric forms of the input molecule are identified and regioselectivity predictions are made for each tautomeric form. The increase in CPU time associated with this capability is offset by pre-screening low energy conformations with the GNFF-xTB force field without significant loss of accuracy. The median CPU time requirements of RegioSQM20 is 49 seconds per molecule on four Intel(R) Xeon(R) CPU X5550 @ 2.67GHz cores, but the computational cost depends heavily on the number of possible reaction sites and number of low energy tautomeric forms. Finally, RegioSQM20 offers a qualitative prediction of the reactivity of each tautomer (low, medium, or high) based on the highest proton affinity, i.e. the energy difference between the unprotonated and most stable protonated form using the most stable tautomer.

RegioSQM20 is developed and tested on 535 molecules for which the regioselectivity of bromination by EAS has been reported in the literature. The accuracy of the predictions with xTB and PM3 are roughly the same (ca. 91%), indicating that the same accuracy can be achieved with an open source approach. The inclusion of tautomers increases the success rate from 90.7% to 92.7%.

RegioSQM20 is compared to two machine learning based models: one developed by Struble et al.[12] specifically for regioselectivity predictions of EAS reactions (WLN) and a more generally applicable reactivity predictor developed by Schwaller et al.[7] (IBM RXN). RegioSQM20 and WLN offers roughly the same success rates for the entire data sets (without considering tautomers), while WLN is many orders of magnitude faster. The accuracy of the more general IBM RXN approach is somewhat lower: 76.3%-85.0%, depending on the data set.

While the WLN based method is considerably faster, the RegioSQM approach may perform better for ring systems that are not well represented in the training set. Unfortunately, the training sets used to develop the WLN based method is not publicly available, so this hypothesis is difficult to check. Another difference is that RegioSQM finds two or more atoms with roughly equal reactivity in about one fourth of the molecules while this almost never happens with the WLN based method. This could reflect the possibility that the literature and especially patents tend to report only the desired product even if other products are observed. This "bias" is then introduced to machine learning models since they are developed based on data that are extracted from these sources. For example, 21 out of the 32 molecules that fail with WLN (without considering tautomers) has more than one experimentally observed reactive site, compared to 6 out of 21 for RegioSQM20*. In fact, out of the 38 molecules in our data set with two or more experimentally observed reaction sites, WLN makes correct prediction for only two molecules, while RegioSQM20* makes correct predictions for 25. RegioSQM20 thus could offer a useful complement to machine learning based methods in some cases.

Acknowledgement

Author's contributions

AG and JHJ developed the idea and lead the project. NR wrote all the code and performed all the calculations. All authors contributed to the data analysis. All authors read and approved the final manuscript.

Funding

This work was supported by Bayer AG

Availability of data and materials

The code is available at https://github.com/jensengroup/RegioSQM20 and data is available at https://github.com/jensengroup/SI_RegioSQM20.

Competing interests

The authors declare that there are no competing interests.

References

(1) Patonay, T.; Kónya, K. *Synthesis and Modification of Heterocycles by Metal-Catalyzed Cross-coupling Reactions (Topics in Heterocyclic Chemistry)*; Springer, 2016.

(2) de Meijere (Editor), A.; (Editor), S. B.; (Editor), M. O. *Metal Catalyzed Cross-Coupling Reactions and More, 3 Volume Set*; Wiley-VCH, 2013.

(3) Segler, M. H. S.; Waller, M. P. *Chemistry - A European Journal* **2017**, *23*, 5966–5971.

(4) Segler, M. H. S.; Preuss, M.; Waller, M. P. *Nature* **2018**, *555*, 604–610.

(5) Coley, C. W. et al. *Science* **2019**, *365*, eaax1566.

(6) Grzybowski, B. A.; Szymkuć, S.; Gajewska, E. P.; Molga, K.; Dittwald, P.; Wołos, A.; Klucznik, T. *Chem* **2018**, *4*, 390–398.

(7) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. *ACS Cent. Sci.* **2019**, *5*, 1572–1583.

(8) Sacha, M.; Blaz, M.; Byrski, P.; Wlodarczyk-Pruszynski, P.; Jastrzebski, S. Molecule Edit Graph Attention Network: Modeling Chemical Reactions as Sequences of Graph Edits. 2020.

(9) Genheden, S.; Thakkar, A.; Chadimova, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. **2020**,

(10) Kruszyk, M.; Jessing, M.; Kristensen, J. L.; Jørgensen, M. *J. Org. Chem.* **2016**, *81*, 5128–5134.

(11) Kromann, J. C.; Jensen, J. H.; Kruszyk, M.; Jessing, M.; Jørgensen, M. *Chem. Sci.* **2018**, *9*, 660–665.

(12) Struble, T. J.; Coley, C. W.; Jensen, K. F. *React. Chem. Eng.* **2020**, *5*, 896–902.

(13) Zubatyuk, R.; Smith, J.; Nebgen, B. T.; Tretiak, S.; Isayev, O. **2020**,

(14) Tomberg, A.; Johansson, M. J.; Norrby, P.-O. *J. Org. Chem.* **2018**, *84*, 4695–4703.

(15) yanfei guan,; Coley, C.; wu, H.; Ranasinghe, D.; esther heid,; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F. **2020**,

(16) Stewart, J. J. P. *Journal of Computational Chemistry* **1989**, *10*, 209–220.

(17) Klamt, A.; Schüürmann, G. *Journal of the Chemical Society, Perkin Transactions 2* **1993**, 799–805.

(18) RDKit: Open-source cheminformatics. `http://www.rdkit.org` (version 2020.03.01).

(19) Spicher, S.; Grimme, S. *Angew. Chem. Int. Ed.* **2020**, *59*, 15665–15673.

(20) Grimme, S.; Bannwarth, C.; Shushkov, P. *J. Chem. Theory Comput.* **2017**, *13*, 1989–2009.

(21) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *J. Cheminf.* **2011**, *3*.

(22) IBM RXN for Chemistry `https://rxn.res.ibm.com` (accessed September 10, 2020).

(23) Streitwieser, A. *Molecular Orbital Theory for Organic Chemists*; John Wiley & Sons Inc, 1961.

Supporting Information
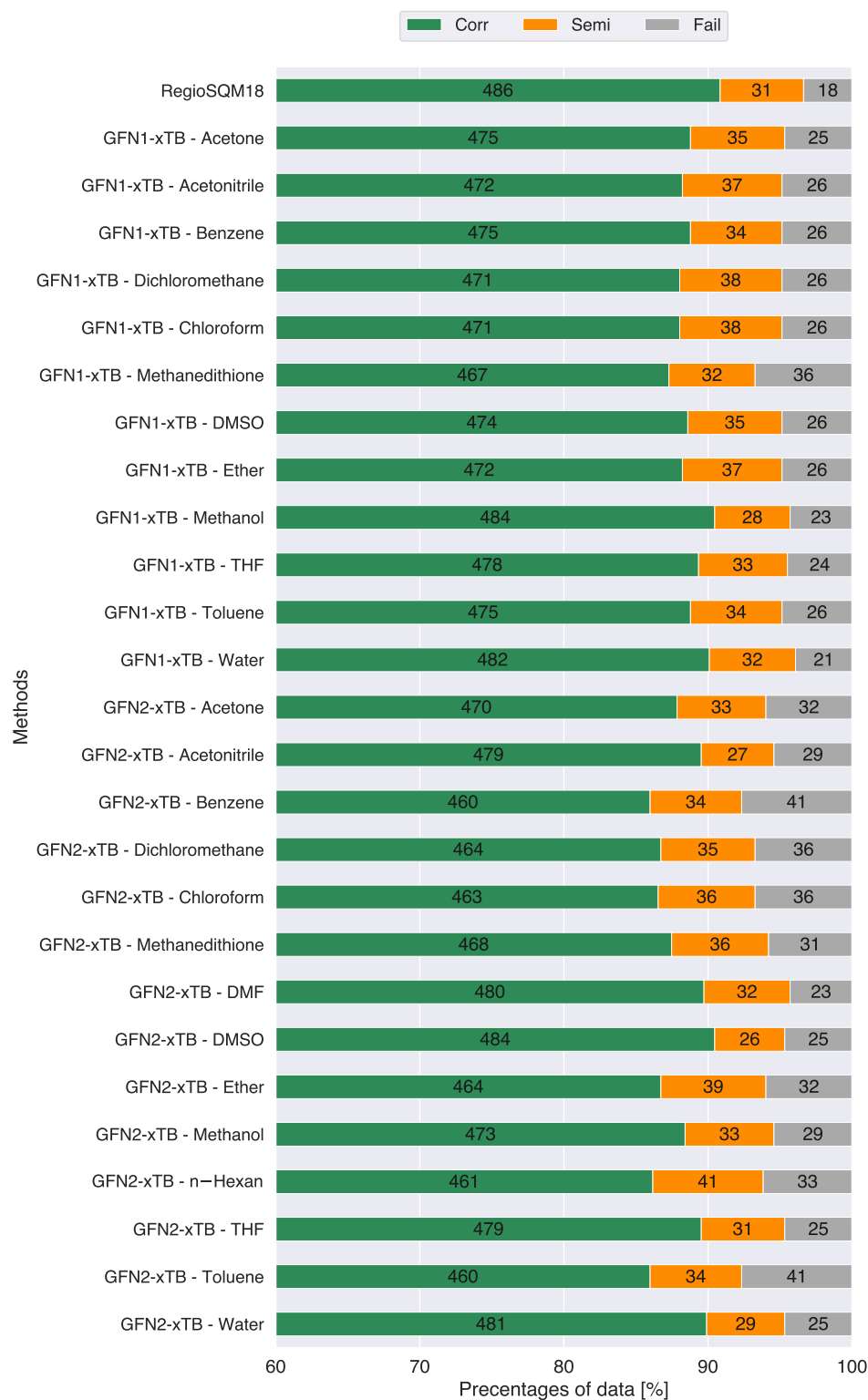
Figure S1: The results of employing different xTB methods in combination with a variety of solvents. All of the starting conformers were generated by RDKit in RegioSQM18 without a random seed, however they were exactly the same for all of the methods shown.
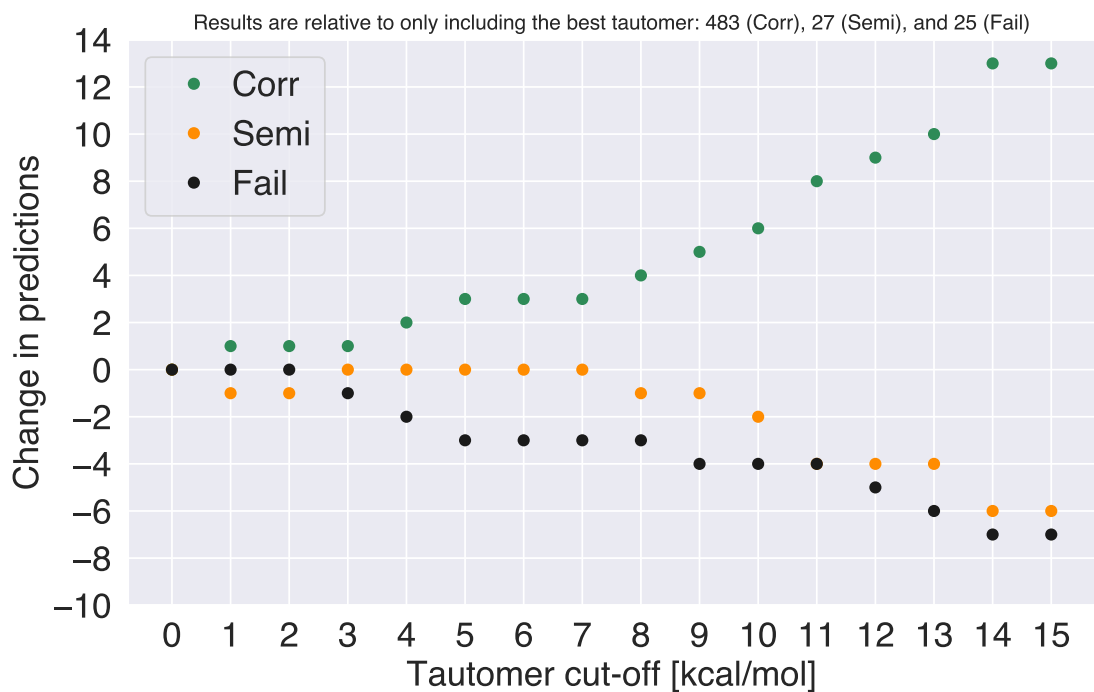
Figure S2: Change in accuracy with increased tautomer cutoff
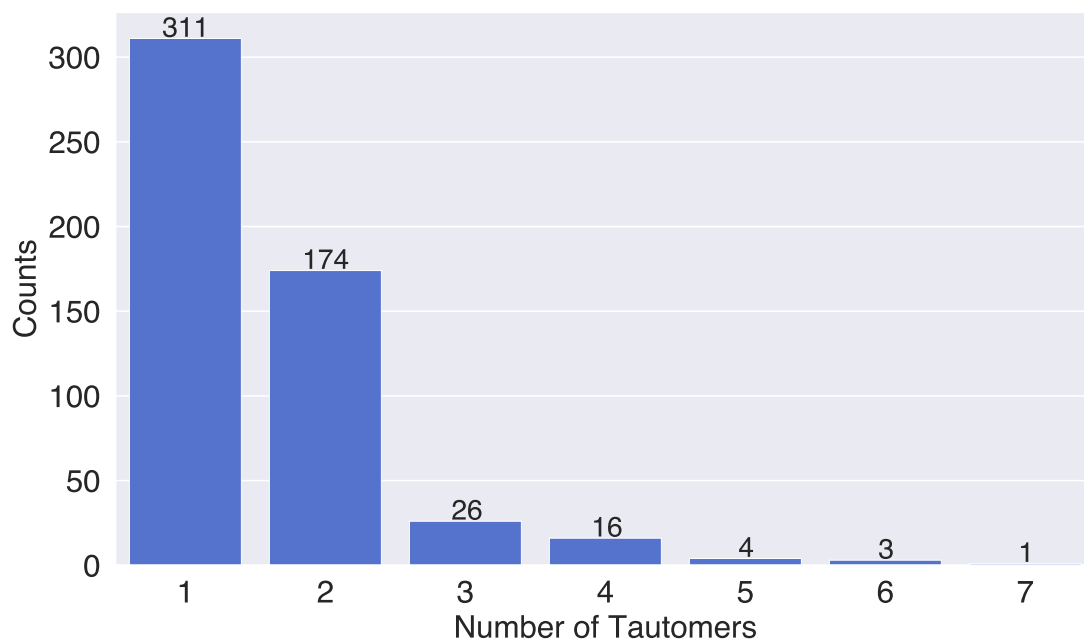
Figure S3: Histogram showing the number of tautomers for each molecule by employing RegioSQM20, which has a tautomer cut-off of 15 kcal/mol.

(a) **73**

(b) **226**

(c) **204**

(d) **382**

(e) **161**

(f) **482**

(g) **323**

(h) **164**

(i) **134**

(j) **339**

(k) **52**

(l) **85**
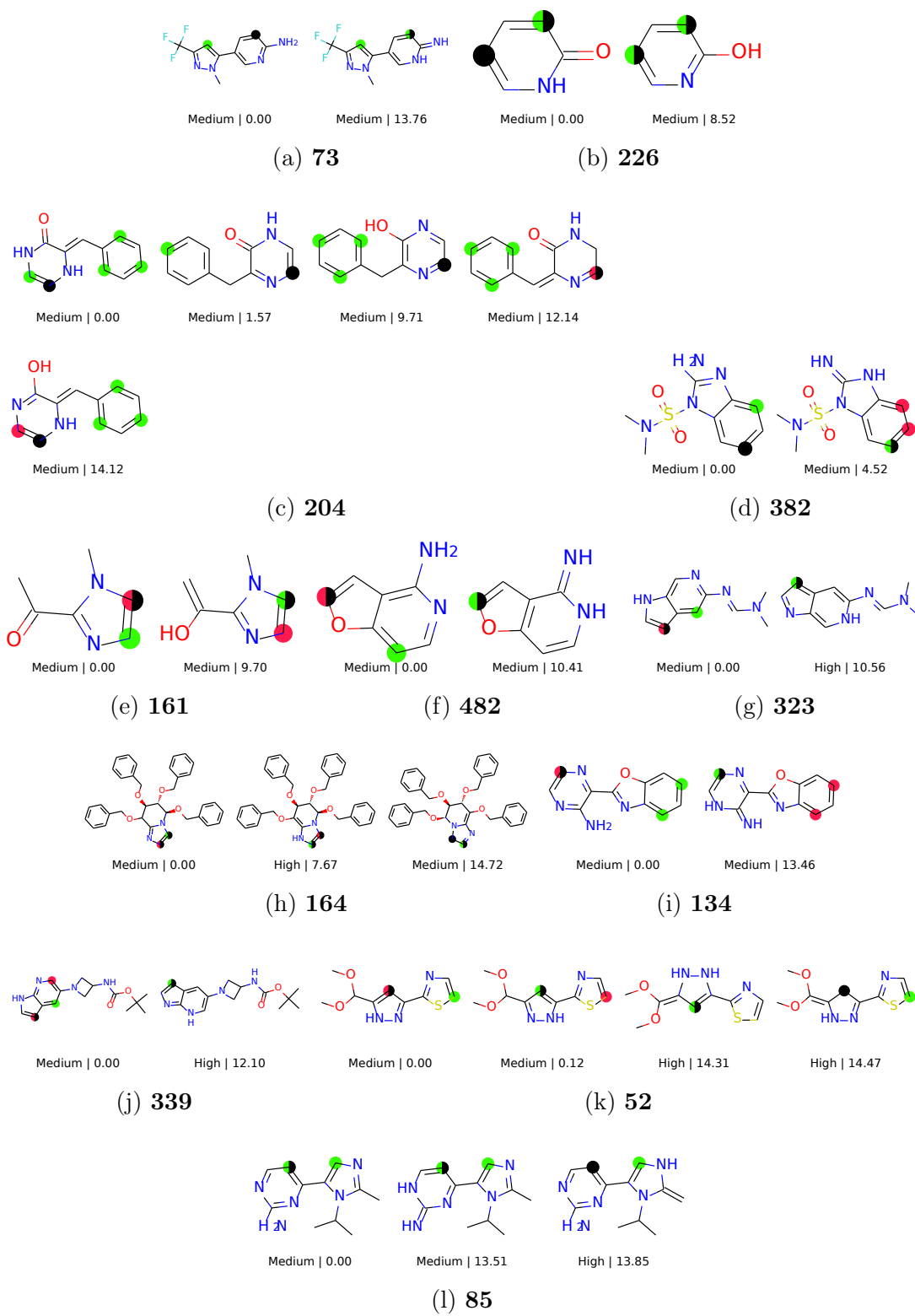
Figure S4: Molecules for which the inclusion of tautomers improves the accuracy.
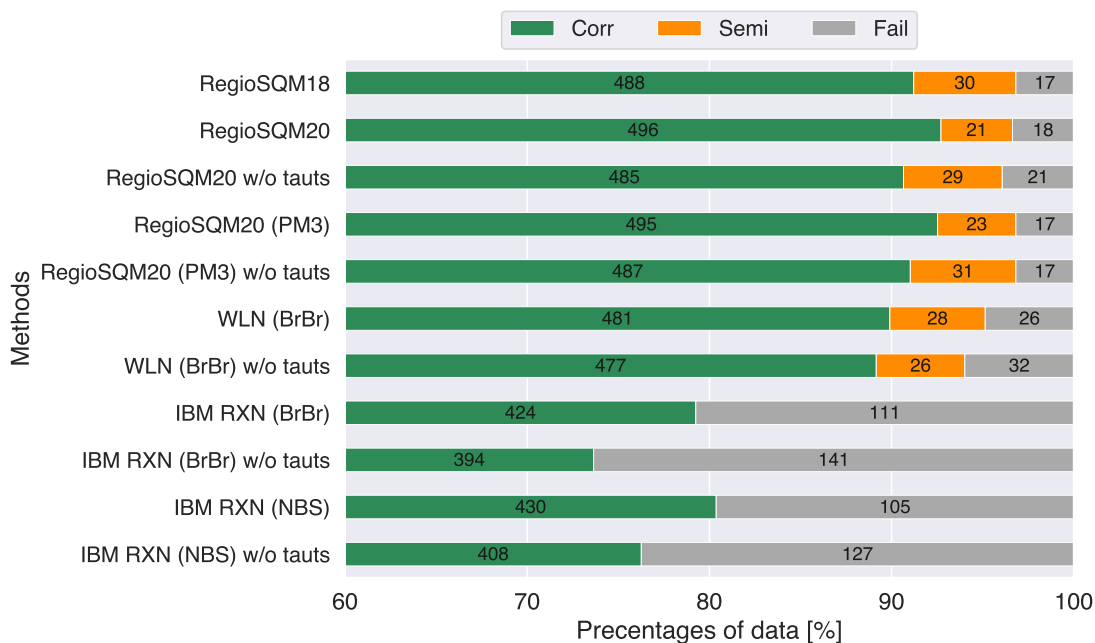
Figure S5: A comparison of different methods used to predict the regioselectivity of the 535 molecules in the data set. "w/o tauts" indicates that the tautomer enumerator is not applied, hence, the results are solely based on the input SMILES. Therefore, "RegioSQM20 w/o tauts" corresponds to "RegioSQM20*" in the paper.
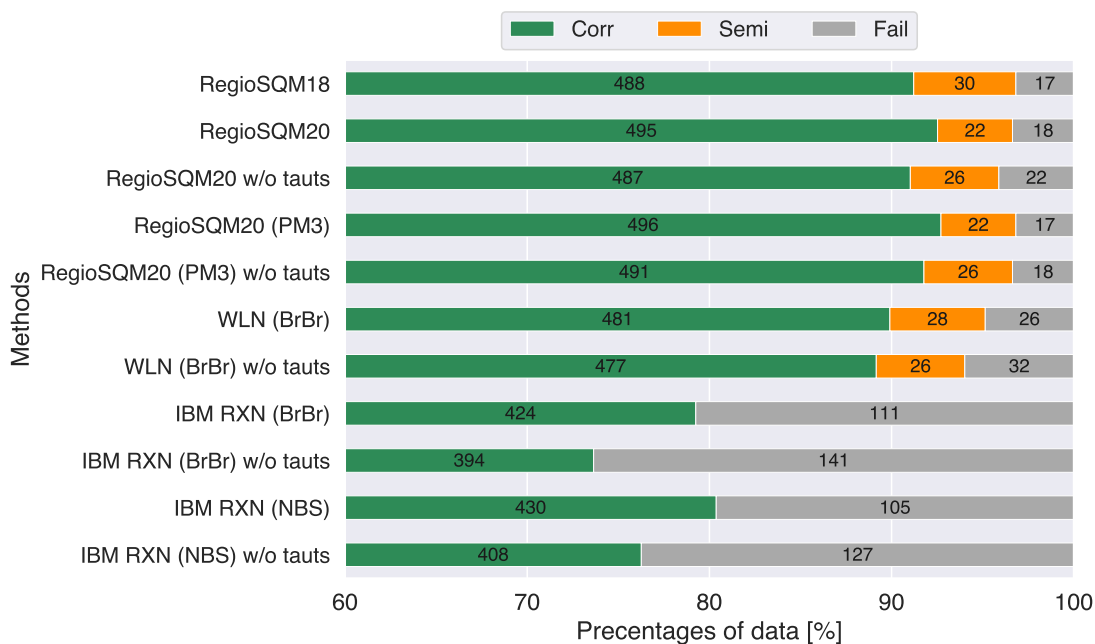
Figure S6: A comparison of different methods without the initial GFNFF-xTB conformer search, which is otherwise used in the four different RegioSQM20 methods.
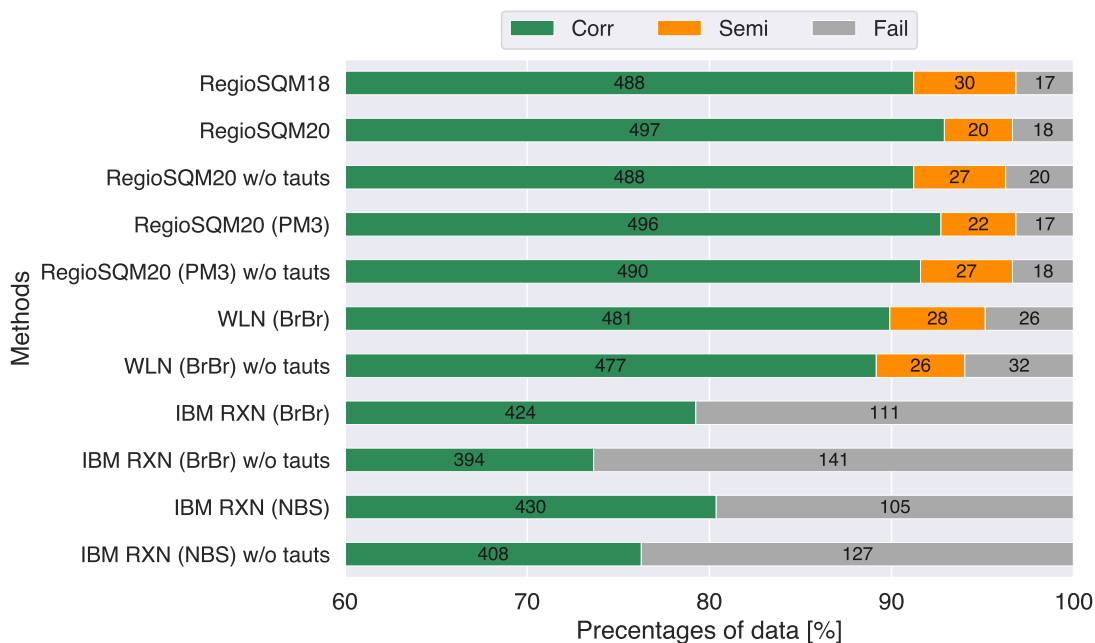
Figure S7: A comparison of different methods with a conformer cutoff of 1000 kcal/mol instead of the default 3 kcal/mol, which is otherwise used in the four different RegioSQM20 methods.

Table S1: The top-1 accuracy of different methods for compounds with one unique black dot[a] and one unique green dot w.r.t. RegioSQM20 w/o tauts (426 out of the 535 compounds in the data set). The results include only the lowest energy tautomer or the tautomer with the overall best atom/confident score.

Name	1 site (%)
RegioSQM20	91.5
RegioSQM20 w/o tauts	92.0
WLN (BrBr)	93.2
WLN (BrBr) w/o tauts	96.0
IBM RXN (BrBr)	81.7
IBM RXN (BrBr) w/o tauts	81.9
IBM RXN (NBS)	84.7
IBM RXN (NBS) w/o tauts	85.0

[a]Out of 535 compounds in the data set, 497 compounds have one unique black dot (experimentally observed reaction site), 37 compounds have two unique black dots, and one compound has three unique black dots.

(a) RegioSQM20
Corr: 483, Semi: 27, Fail 25

(b) RegioSQM20 w/o tauts
Corr: 485, Semi: 29, Fail 21
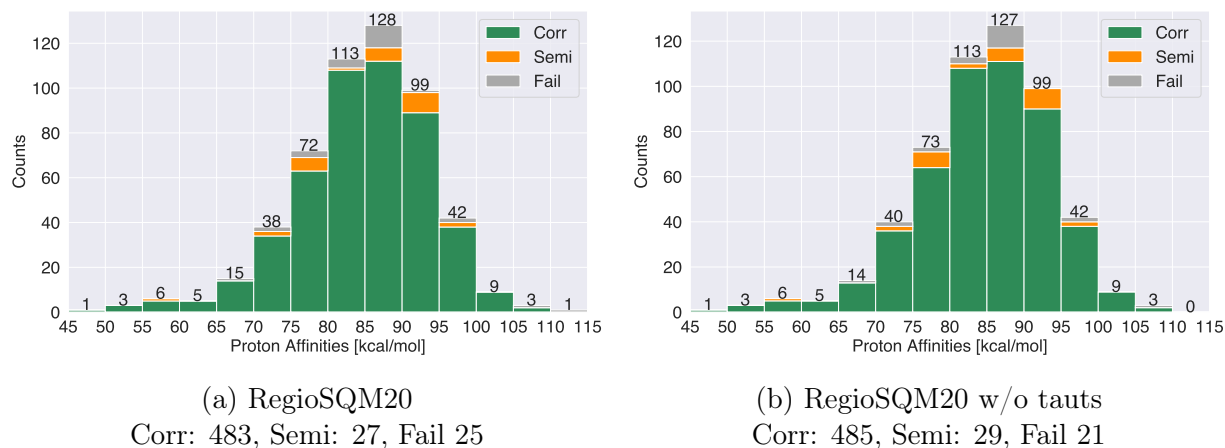
Figure S8: Histograms of the highest proton affinities for the 535 molecules in the data set using (a) RegioSQM20 and (b) RegioSQM20 w/o tauts. The values are given in units of kcal/mol. For the RegioSQM20 results, only the lowest energy tautomer of each molecule is considered.
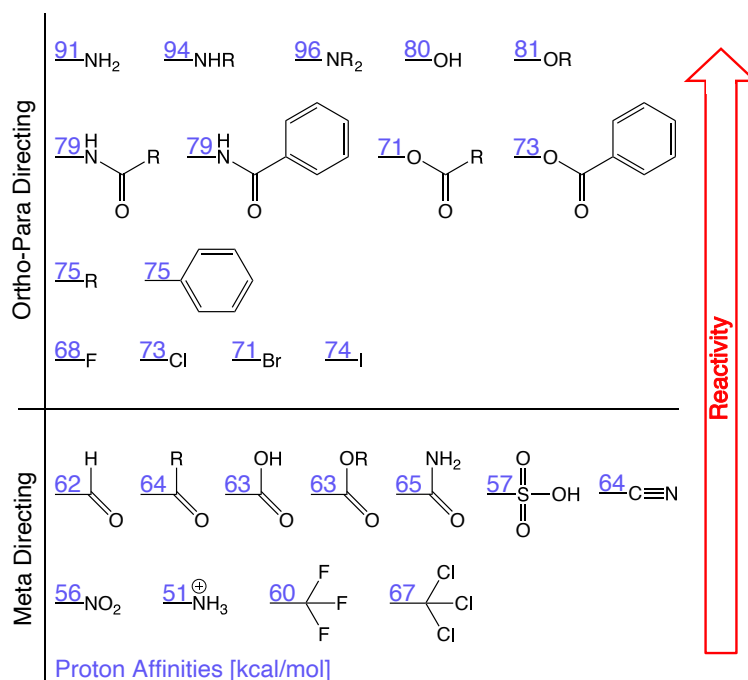
Figure S9: The highest proton affinities computed for a series of substituted benzene analogs using RegioSQM20. The values are given in units of kcal/mol and only the lowest energy tautomer of each molecule is considered.

20

(a) RegioSQM20 (PM3)
Corr: 485, Semi: 32, Fail 18

(b) RegioSQM20 (PM3) w/o tauts
Corr: 487, Semi: 31, Fail 17
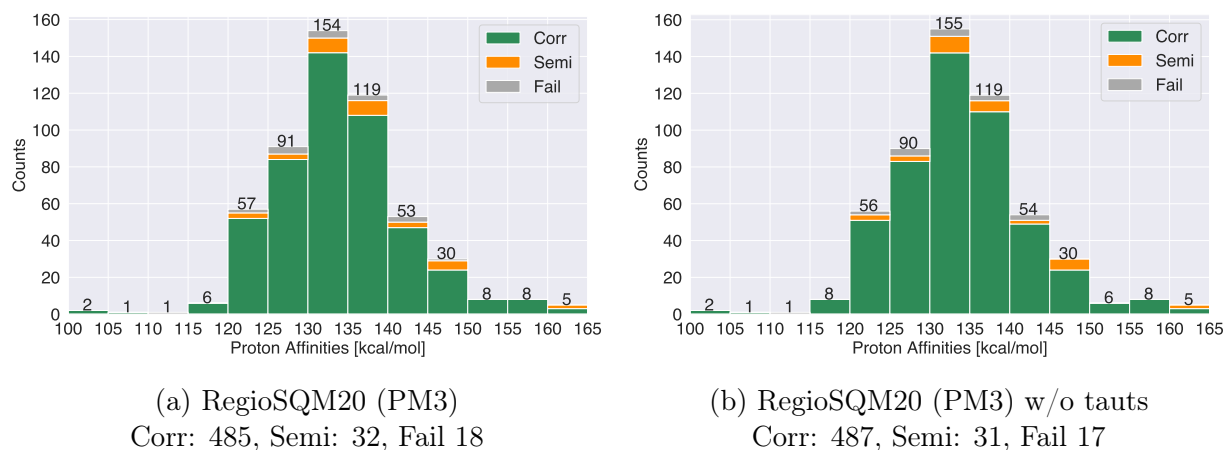
Figure S10: Histograms showing the difference in heat of formation between the lowest energy protonated form and the unprotonated form of each molecule in the data set using (a) RegioSQM20 (PM3) and (b) RegioSQM20 (PM3) w/o tauts. The energies are given in units of kcal/mol. For the RegioSQM20 (PM3) results, only the lowest energy tautomer of each molecule is considered.
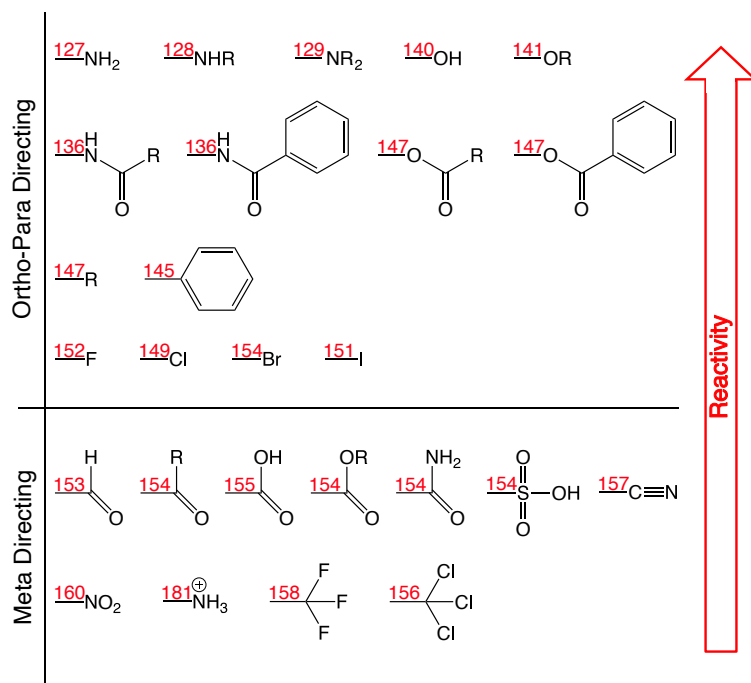
Figure S11: The difference in heat of formation between the lowest energy protonated form and the unprotonated form of each molecule in a series of substituted benzene analogs using RegioSQM20 (PM3). The energies are given in units of kcal/mol and only the lowest energy tautomer of each molecule is considered. Note, the reactivity increases with lower values which is the opposite of Fig. S9, since the heat of formation of the proton is neglected in the values presented here.

21
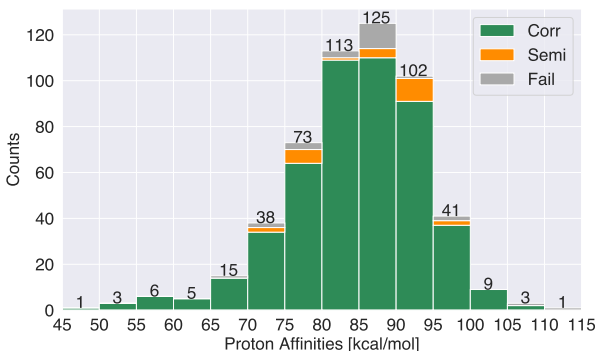
(a) RegioSQM20
Corr: 485, Semi: 25, Fail 25
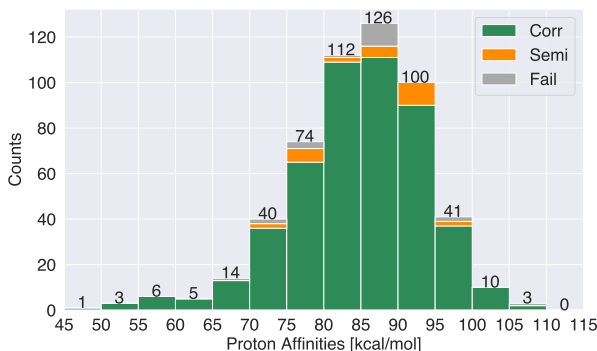
(b) RegioSQM20 w/o tauts
Corr: 488, Semi: 27, Fail 20

Figure S12: Histograms of the highest proton affinities using (a) RegioSQM20 and (b) RegioSQM20 w/o tauts with a conformer cutoff of 1000 kcal/mol instead of the default 3 kcal/mol. For the RegioSQM20 results, only the lowest energy tautomer of each molecule is considered.
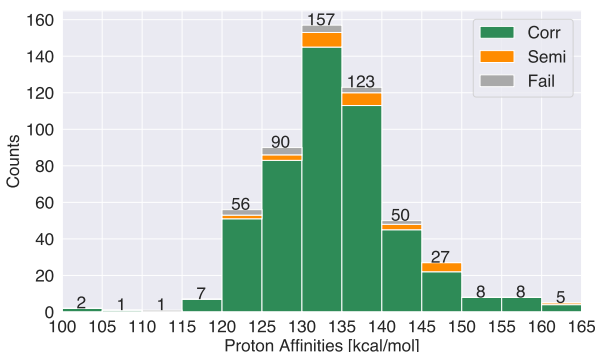
(a) RegioSQM20 (PM3)
Corr: 489, Semi: 29, Fail 17
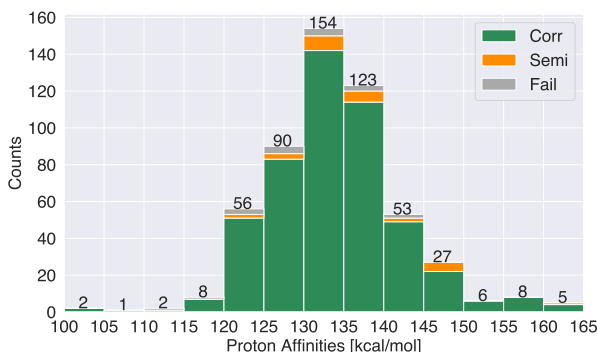
(b) RegioSQM20 (PM3) w/o tauts
Corr: 490, Semi: 27, Fail 18

Figure S13: Histograms showing the difference in heat of formation between the lowest energy protonated form and the unprotonated form of each molecule in the data set using (a) RegioSQM20 (PM3) and (b) RegioSQM20 (PM3) w/o tauts with a conformer cutoff of 1000 kcal/mol instead of the default 3 kcal/mol. The energies are given in units of kcal/mol. For the RegioSQM20 (PM3) results, only the lowest energy tautomer of each molecule is considered.

22
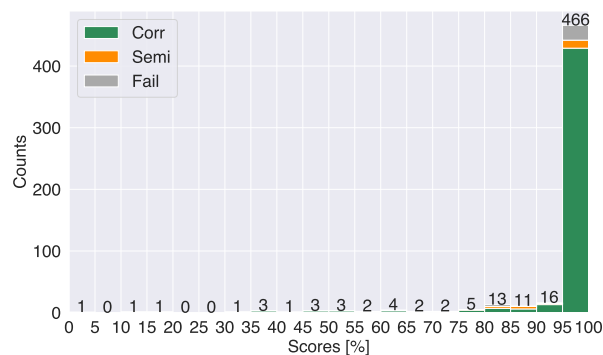
(a) WLN (BrBr) w/o tauts
Corr: 477, Semi: 26, Fail 32
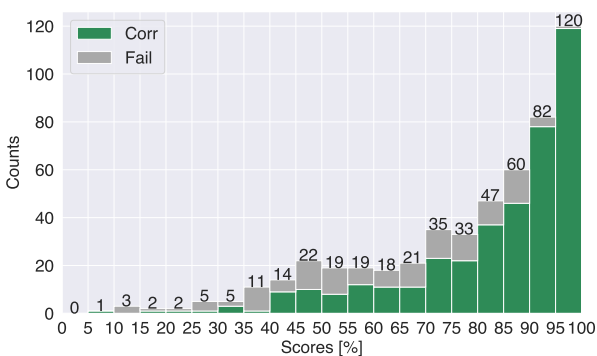
(b) IBM RXN (BrBr) w/o tauts
Corr: 394, Fail 125 (not completed: 16)
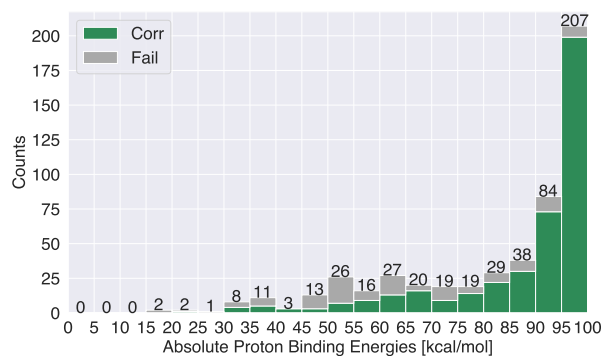
(c) IBM RXN (NBS) w/o tauts
Corr: 408, Fail 117 (not completed: 10)

Figure S14: Histograms showing (a) the best atom score using WLN (BrBr), (b) the confident score using IBM RXN (BrBr), and (c) the confident score using IBM RXN (NBS). All results are solely based on the input SMILES and do not include different tautomers.
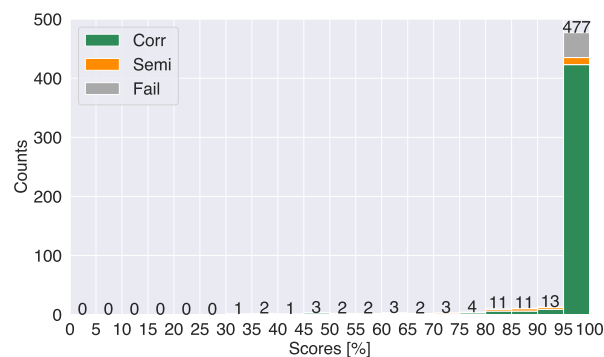
(a) WLN (BrBr).
Corr: 460, Semi: 27, Fail 48
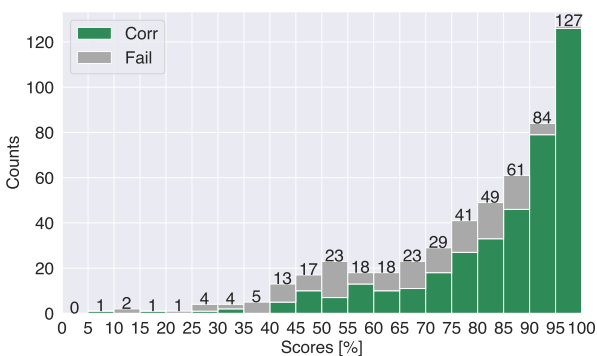
(b) IBM RXN (BrBr).
Corr: 390, Fail 131 (not completed: 14)
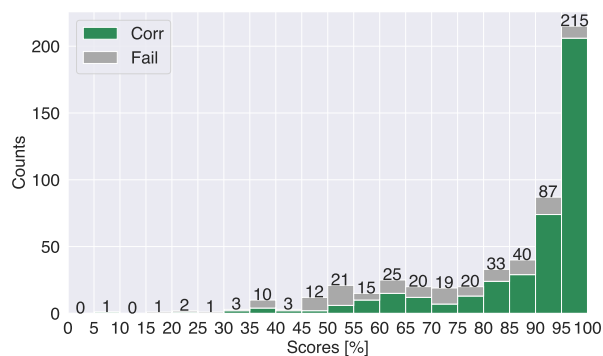
(c) IBM RXN (NBS).
Corr: 408, Fail 120 (not completed: 7)

Figure S15: Histograms showing (a) the best atom score among the tautomers using WLN (BrBr), (b) the best confident score among the tautomers using IBM RXN (BrBr), and (c) the best confident score among the tautomers using IBM RXN (NBS).
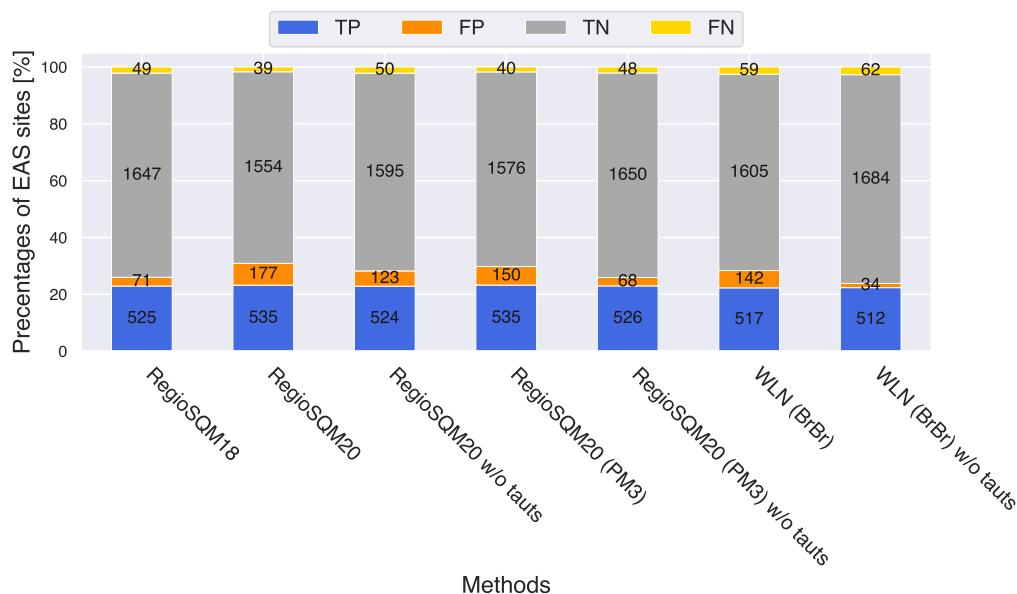
Figure S16: Confusion matrix of all possible reaction sites in the data set. A positive result corresponds to a reaction site being classified as **"Corr"** for any of the tautomers below 15 kcal/mol w.r.t. the RegioSQM methods or all of the generated tautomers w.r.t. the WLN methods. Only unique reaction centers are taken into account.
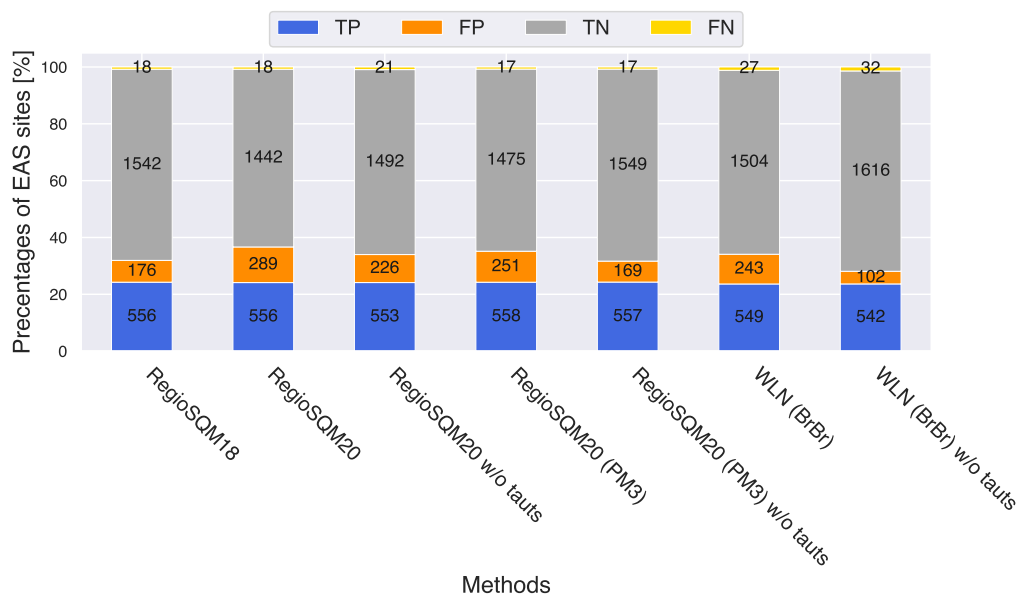
Figure S17: Confusion matrix of all possible reaction sites in the data set. A positive result corresponds to a reaction site being classified as **"Semi"** for any of the tautomers below 15 kcal/mol w.r.t. the RegioSQM methods or all of the generated tautomers w.r.t. the WLN methods. Only unique reaction centers are taken into account.
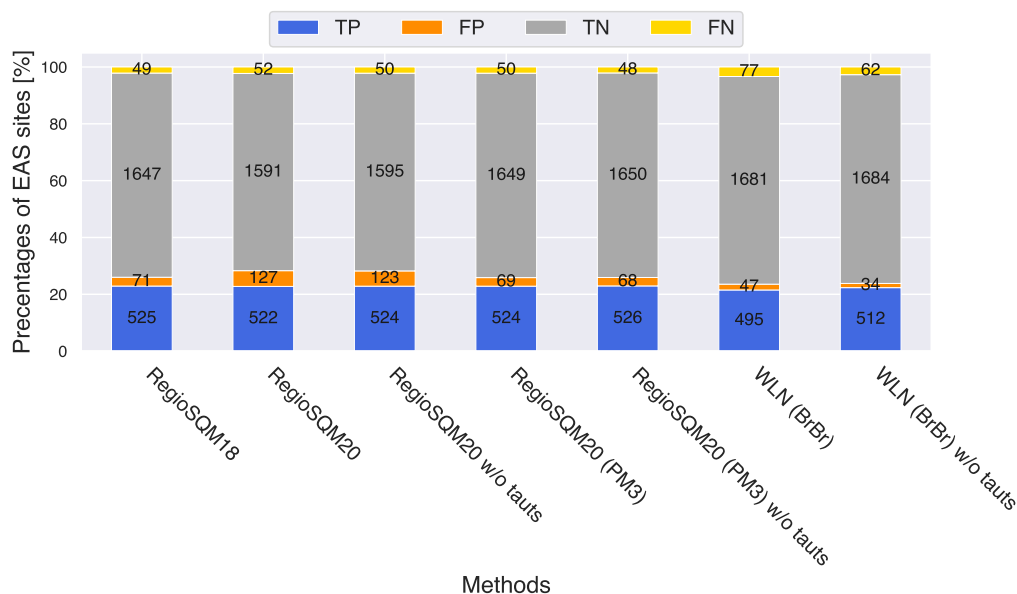
Figure S18: Confusion matrix of all possible reaction sites in the data set. A positive result corresponds to a reaction site being classified as **"Corr"** for only the lowest energy unprotonated tautomer w.r.t. the RegioSQM methods or the tautomer with the overall best atom score w.r.t. the WLN methods. Only unique reaction centers are taken into account.
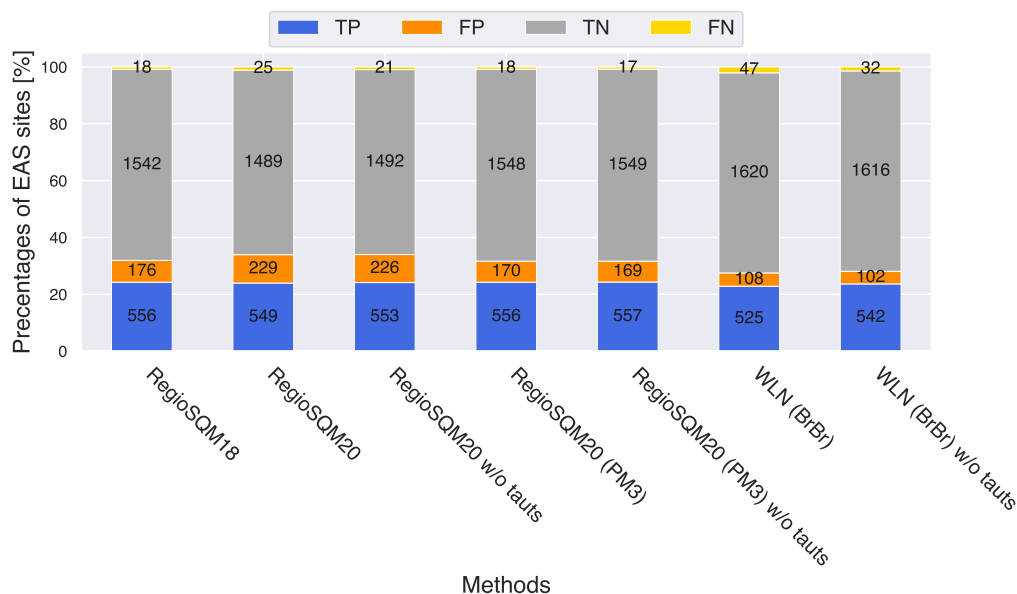
Figure S19: Confusion matrix of all possible reaction sites in the data set. A positive result corresponds to a reaction site being classified as **"Semi"** for only the lowest energy unprotonated tautomer w.r.t. the RegioSQM methods or the tautomer with the overall best atom score w.r.t. the WLN methods. Only unique reaction centers are taken into account.

Results are relative to only including the best tautomer: 522 (TP), 127 (FP), 1591 (TN), 52 (FN)
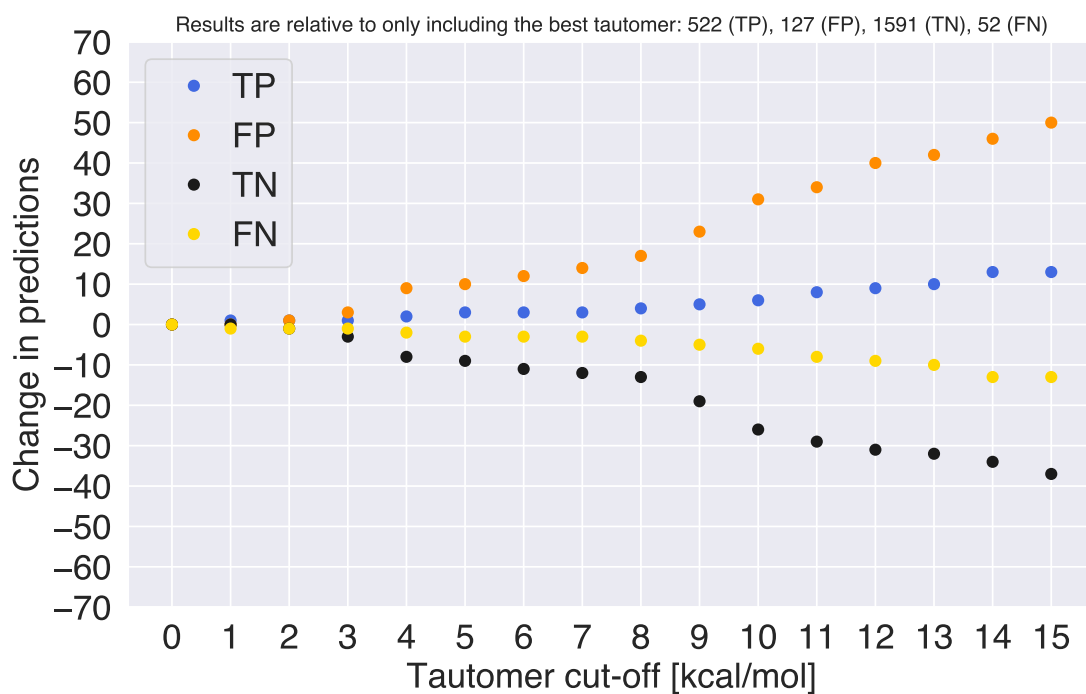
Figure S20: RegioSQM20. Confusion matrix - Corr

Results are relative to only including the best tautomer: 549 (TP), 229 (FP), 1489 (TN), 25 (FN)
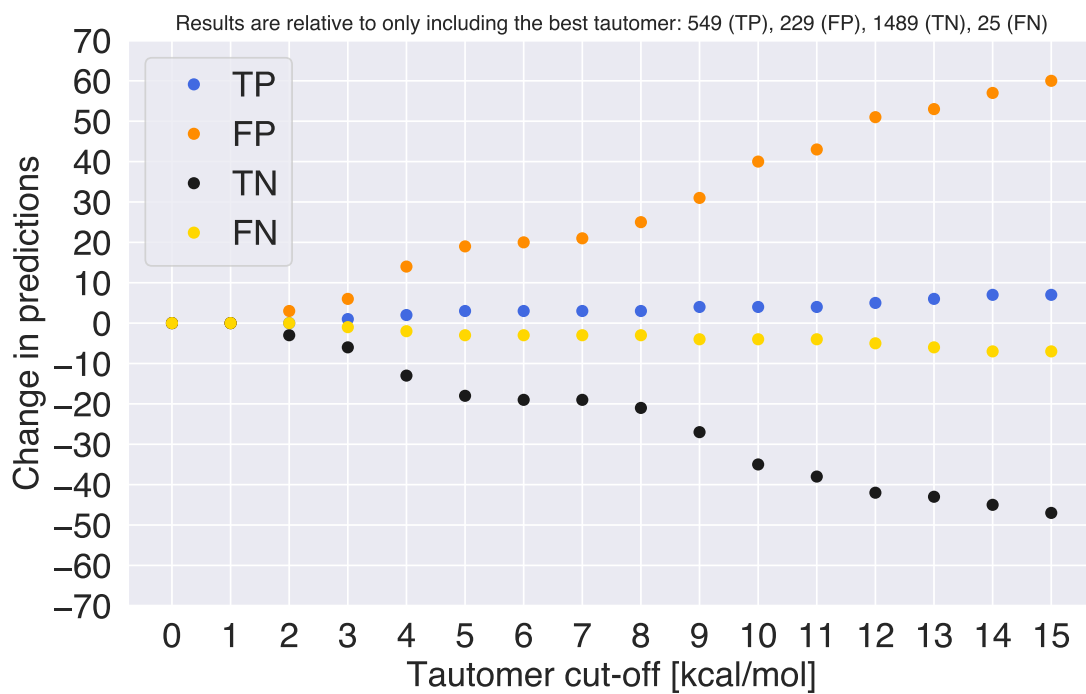
Figure S21: RegioSQM20. Confusion matrix - Semi

Predictions using RegioSQM20
for the 535 molecules in the data set

Table of Contents

Pyrroles

Medium | 0.00

(S22.1) **402**

Furans

Medium | 0.00

(S23.1) **434**

Medium | 0.00

(S23.2) **435**

Medium | 0.00

(S23.3) **437**

Medium | 0.00

(S23.4) **438**

Medium | 0.00

(S23.5) **441**

Medium | 0.00

(S23.6) **442**

Medium | 0.00

Medium | 5.33

(S23.7) **444**

Medium | 0.00

(S23.8) **446**

Medium | 0.00

(S23.9) **449**

Medium | 0.00

(S23.10) **450**

Medium | 0.00

Medium | 0.00

Medium | 0.00

Medium | 0.00

Medium | 14.99

Medium | 14.99

Medium | 14.99

Medium | 14.99

Medium | 0.00

(S23.11) **451**

(S23.12) **452**

Medium | 0.00

Medium | 0.00

Medium | 12.28

Medium | 12.93

Medium | 13.49

High | 14.70

High | 14.70

(S23.13) **453**

Medium | 0.00

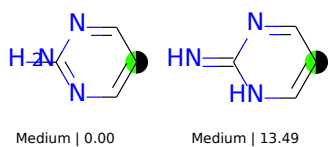(S23.14) **454**

Medium | 0.00

(S23.15) **465**

Medium | 0.00

(S23.16) **466**

Medium | 8.40

Medium | 0.00

(S23.17) **469**

Medium | 0.00

(S23.18) **474**

Medium | 0.00

(S23.19) **476**

Medium | 0.00

(S23.20) **477**

Medium | 0.00

(S23.21) **478**

Medium | 0.00

(S23.22) **479**

Medium | 0.00

(S23.23) **483**

Medium | 0.00

(S23.24) **484**

High | 0.00   Medium | 13.94

(S23.25) **485**

Medium | 0.00   Medium | 13.76

(S23.26) **486**

Medium | 0.00

(S23.27) **443**

Medium | 0.00

(S23.28) **475**

Thiophenes

Medium | 0.00

(S24.1) **413**

Medium | 0.00   Medium | 11.52

(S24.2) **586**

Medium | 0.00   Medium | 11.72

(S24.3) **588**

Medium | 0.00   Medium | 13.05

(S24.4) **587**

Pyrazoles

Medium | 0.00

(S25.1) **1**

Medium | 0.00

(S25.2) **2**

High | 0.00

(S25.3) **3**

Medium | 0.00

(S25.4) **4**

Medium | 0.00   Medium | 1.73

(S25.5) **5**

Medium | 0.00   Medium | 1.67

(S25.6) **50**

Medium | 0.00   Medium | 0.09

(S25.7) **51**

Medium | 0.00   Medium | 0.12   High | 14.31   High | 14.47

(S25.8) **52**

Medium | 0.00

(S25.9) **53**

Medium | 0.00   High | 13.29

(S25.10) **54**

Medium | 0.00   Medium | 14.80

(S25.11) **55**

Medium | 0.00   Medium | 0.66

(S25.12) **56**

Medium | 0.00

(S25.13) **216**

Medium | 0.00

(S25.14) **278**

Low | 0.00

(S25.15) **392**

Medium | 0.00

(S25.16) **397**

Medium | 0.00

(S25.17) **421**

High | 0.00

(S25.18) **599**

High | 0.00

(S25.19) **600**

High | 0.00

(S25.20) **601**

Imidazoles

Medium | 0.00

(S26.1) **14**

Medium | 0.47

(S26.2) **15**

Medium | 0.00

(S26.3) **16**

Medium | 0.00

(S26.4) **17**

Medium | 0.00

(S26.5) **58**

Medium | 0.00

(S26.6) **125**

Medium | 0.00

(S26.7) **126**

Medium | 0.00

(S26.8) **127**

Medium | 0.00

(S26.9) **155**

Medium | 0.00

(S26.10) **156**

Medium | 0.00

(S26.11) **157**

Medium | 0.00

(S26.12) **158**

Low | 0.00

(S26.13) **160**

Medium | 0.00

(S26.14) **161**

Medium | 9.70

Medium | 0.00

(S26.15) **162**

Medium | 0.00

(S26.16) **165**

Medium | 0.00

(S26.17) **166**

Medium | 0.00

(S26.18) **171**

Medium | 0.00

(S26.19) **173**

Medium | 0.00

(S26.20) **137**

Medium | 0.00

(S26.21) **154**

Medium | 0.00

(S26.22) **159**

Medium | 0.00

(S26.23) **163**

Medium | 0.00

High | 7.67

164

Medium | 14.72

(S26.24) **164**

Medium | 0.00

(S26.25) **168**

Medium | 0.00

Medium | 9.05

(S26.26) **169**

Medium | 0.00

(S26.27) **401**

Isoxazoles

Medium | 0.00

(S27.1) **6**

Medium | 0.00

(S27.2) **7**

Medium | 0.00

(S27.3) **8**

Medium | 0.00

(S27.4) **9**

Medium | 0.00

(S27.5) **141**

Oxazoles

Medium | 0.00

(S28.1) **43**

Medium | 0.00

(S28.2) **119**

Medium | 0.00

(S28.3) **122**

Isothiazoles

Medium | 0.00

(S29.1) **10**

Medium | 0.00

(S29.2) **11**

Medium | 0.00

(S29.3) **13**

Medium | 0.00

(S29.4) **598**

Thiazoles

Medium | 0.00

Medium | 0.00

Medium | 0.00

Medium | 0.00

Medium | 0.00

(S30.1) **18** (S30.2) **19** (S30.3) **20** (S30.4) **59** (S30.5) **60**

Medium | 0.00

Medium | 4.65

Medium | 0.00

Medium | 0.00

(S30.6) **61** (S30.7) **62** (S30.8) **64**

Medium | 0.00

Medium | 1.74

High | 5.22

Medium | 0.00

Medium | 11.65

(S30.9) **66** (S30.10) **67**

1,2,4-Triazoles

Low | 0.00

Low | 0.92

Medium | 1.52

(S31.1) **147**

Medium | 0.00

Medium | 1.72

Medium | 2.28

Medium | 7.57

Medium | 9.00

Medium | 9.32

Medium | 14.48

(S31.2) **149**

38

Benzenes

Medium | 0.00

(S32.1) **274**

Medium | 0.00   Medium | 3.91   Medium | 4.24   Medium | 11.76

(S32.2) **276**

Medium | 0.00   High | 6.54

(S32.3) **277**

Low | 0.00

(S32.4) **428**

Low | 0.00

(S32.5) **488**

Low | 0.00

(S32.6) **489**

Medium | 0.00

(S32.7) **490**

Low | 0.00

(S32.8) **493**

Low | 0.00

(S32.9) **494**

Low | 0.00

(S32.10) **495**

Low | 0.00

(S32.11) **496**

Low | 0.00

(S32.12) **497**

Medium | 0.00

(S32.13) **498**

Low | 0.00

(S32.14) **499**

Medium | 0.00

(S32.15) **500**

Medium | 0.00

(S32.16) **501**

Medium | 0.00

(S32.17) **502**

Medium | 0.00

(S32.18) **503**

Medium | 0.00

(S32.19) **504**

Medium | 0.00

(S32.20) **509**

Medium | 0.00

(S32.21) **511**

Medium | 0.00

(S32.22) **514**

Medium | 0.00

(S32.23) **515**

Low | 0.00

(S32.24) **516**

Low | 0.00

(S32.25) **517**

Medium | 0.00

(S32.26) **518**

Medium | 0.00

(S32.27) **519**

Low | 0.00

(S32.28) **520**

Medium | 0.00

(S32.29) **521**

Low | 0.00   Medium | 7.67

(S32.30) **522**

Medium | 0.00

(S32.31) **523**

Medium | 0.00

(S32.32) **526**

Medium | 0.00

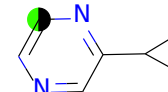(S32.33) **529**

Low | 0.00

(S32.34) **530**

Medium | 0.00

(S32.35) **531**

Medium | 0.00

(S32.36) **532**

Medium | 0.00

(S32.37) **534**

Medium | 0.00

(S32.38) **535**

Medium | 0.00

(S32.39) **536**

Medium | 0.00

(S32.40) **537**

Medium | 0.00

(S32.41) **539**

Medium | 0.00

(S32.42) **540**

Medium | 0.00

(S32.43) **541**

Medium | 0.00

(S32.44) **542**

Medium | 0.00

(S32.45) **543**

Medium | 10.69

(S32.46) **546**

Medium | 0.00

(S32.47) **549**

Medium | 0.00

(S32.48) **550**

Medium | 0.00

(S32.49) **551**

High | 9.13

(S32.50) **552**

Medium | 0.00

(S32.51) **553**

Medium | 0.00

(S32.52) **554**

Medium | 0.00

(S32.53) **555**

Medium | 0.00

(S32.54) **556**

Medium | 0.00

(S32.55) **558**

High | 0.00

(S32.56) **559**

Medium | 0.00

(S32.57) **560**

Medium | 0.00

(S32.58) **562**

Medium | 0.00

(S32.59) **563**

Medium | 0.00

(S32.60) **564**

Medium | 0.00

(S32.61) **566**

Medium | 0.00

(S32.62) **567**

Medium | 0.00

(S32.63) **568**

Medium | 0.00

High | 11.00

(S32.64) **569**

(S32.65) **570**

Medium | 0.00

(S32.66) **571**

Medium | 0.00

High | 14.35

(S32.67) **572**

Medium | 0.00

(S32.68) **573**

Medium | 0.00

(S32.69) **575**
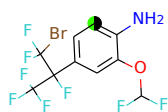
Medium | 0.00

(S32.70) **576**
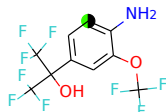
Medium | 0.00

(S32.71) **577**
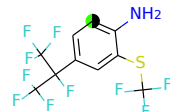
Medium | 0.00

(S32.72) **578**
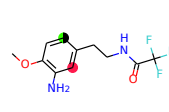
Medium | 0.00

(S32.73) **579**
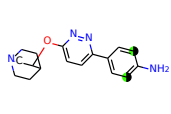
Medium | 0.00

(S32.74) **582**
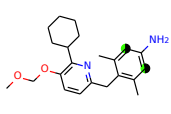
Medium | 0.00

(S32.75) **583**
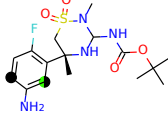
Medium | 0.00

(S32.76) **584**
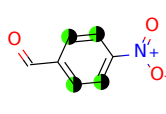
Low | 0.00

(S32.77) **492**
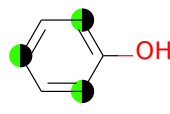
Medium | 0.00

(S32.78) **505**
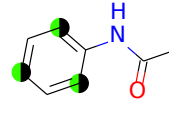
Medium | 0.00

(S32.79) **524**
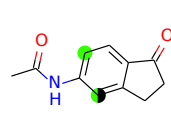
Medium | 0.00

(S32.80) **528**
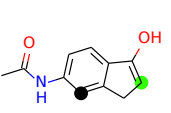
High | 8.32

Medium | 0.00

(S32.81) **545**
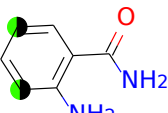
Medium | 0.00

(S32.82) **547**
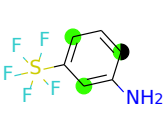
Medium | 0.00

(S32.83) **561**
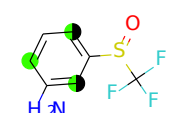
Medium | 0.00

(S32.84) **565**
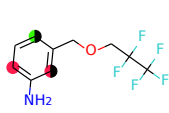
Medium | 0.00

(S32.85) **508**
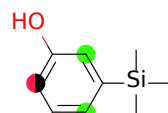
Low | 0.00

(S32.86) **533**

Pyridines

Medium | 0.00
(S33.1) **21**

Medium | 0.00
Medium | 14.10
(S33.2) **22**

Medium | 0.00
Medium | 13.13
(S33.3) **23**

Medium | 0.00
Medium | 9.15
(S33.4) **24**

Medium | 0.00
(S33.5) **48**

Medium | 0.00
Medium | 13.21
(S33.6) **68**

Medium | 0.00
Medium | 13.38
(S33.7) **69**

Medium | 0.00
(S33.8) **70**

Medium | 0.00
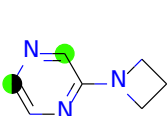High | 7.80
Medium | 12.49
(S33.9) **71**

Medium | 0.00
Medium | 13.76
(S33.10) **73**

Medium | 0.00
Medium | 9.73
(S33.11) **75**

Medium | 0.00
Medium | 13.38
(S33.12) **76**

Medium | 0.00
Medium | 11.10
(S33.13) **124**

Medium | 0.00
(S33.14) **128**

Medium | 0.00
(S33.15) **129**

Medium | 0.00
(S33.16) **130**

Medium | 0.00
(S33.17) **215**

Medium | 0.00
(S33.18) **216**

Medium | 0.00
(S33.19) **217**

Medium | 13.47
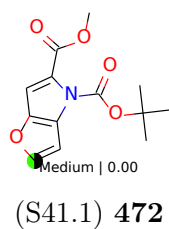
Medium | 0.00
(S33.20) **218**

Medium | 0.00
Medium | 14.01
(S33.21) **221**

Medium | 0.00
(S33.22) **223**

Medium | 0.00
Medium | 13.55
(S33.23) **224**

Medium | 0.00
Medium | 3.95
(S33.24) **225**

Medium | 0.00
(S33.25) **228**

Medium | 0.00

(S33.26) **229**

Medium | 0.00

(S33.27) **234**

High | 13.34

(S33.28) **236**

Medium | 0.00

(S33.29) **237**

Medium | 0.00

(S33.30) **238**

Medium | 0.00

(S33.31) **239**

Medium | 0.00

(S33.32) **241**

Medium | 7.95

(S33.33) **242**

Medium | 0.00

(S33.34) **243**

Medium | 0.00

Low | 0.00

(S33.35) **244**

Medium | 0.00

(S33.36) **247**

Medium | 0.00

Medium | 10.55

(S33.37) **248**

Medium | 0.00

Medium | 14.90

(S33.38) **249**

Medium | 0.00

(S33.39) **250**

Medium | 0.00

(S33.40) **252**

Medium | 0.00

(S33.41) **253**

Medium | 0.00

(S33.42) **255**

Medium | 0.00

Medium | 13.22

(S33.43) **256**
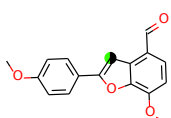
Medium | 0.00

(S33.44) **257**

Medium | 0.00

(S33.45) **258**

Medium | 0.00

Medium | 13.05

(S33.46) **259**

Medium | 0.00

(S33.47) **296**

Medium | 0.00

(S33.48) **297**

Medium | 0.00

Medium | 7.21

(S33.49) **298**

Medium | 0.00

Medium | 13.92

(S33.50) **391**

Medium | 0.00

Medium | 12.64

(S33.51) **395**

Medium | 0.00

(S33.52) **400**

Medium | 13.70

Medium | 0.00

(S33.53) **420**

Medium | 12.66

Medium | 0.00

(S33.54) **74**

Medium | 0.00

(S33.55) **219**

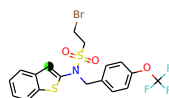Medium | 0.00

(S33.56) **220**

Medium | 0.00

(S33.57) **235**

High | 13.37

Medium | 0.00
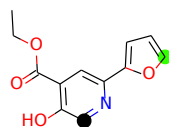
(S33.58) **245**
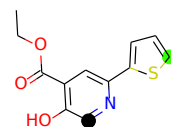
Medium | 0.00

(S33.59) **246**
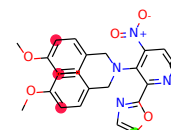
Medium | 0.00

(S33.60) **251**
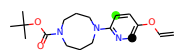
Medium | 0.00

(S33.61) **254**

44

2-Pyridones
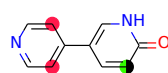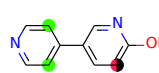
Medium | 0.00   Medium | 8.05   Medium | 13.85   Medium | 0.00   Medium | 8.35
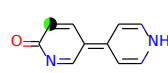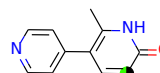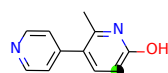
(S34.1) **77**   (S34.2) **78**
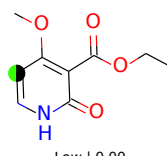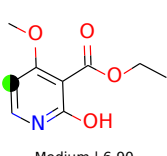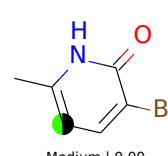
Low | 0.00   Medium | 6.90   Medium | 0.00   Medium | 8.94   Medium | 0.00
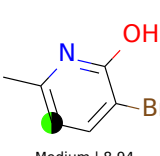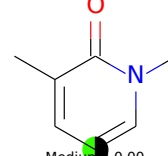
(S34.3) **79**   (S34.4) **80**   (S34.5) **81**
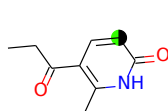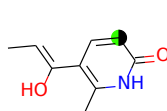
Medium | 0.00   Medium | 8.83   Medium | 8.96   Medium | 0.00   Low | 7.54
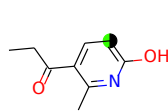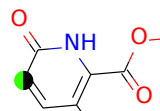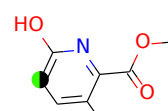
(S34.6) **82**   (S34.7) **83**
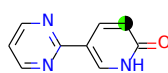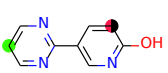
Medium | 0.00   Medium | 8.10   Medium | 0.00   Medium | 9.28   Medium | 0.00   Medium | 9.07
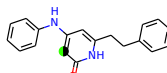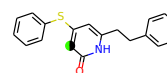
(S34.8) **84**   (S34.9) **213**   (S34.10) **214**
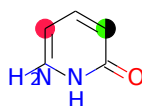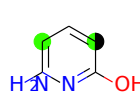
Medium | 0.00   Medium | 8.08   Medium | 0.00   Medium | 9.70   Medium | 0.00   Medium | 10.49
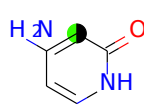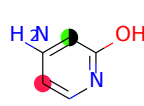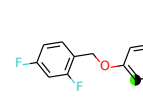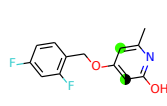
(S34.11) **227**   (S34.12) **230**   (S34.13) **233**
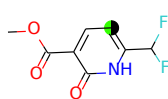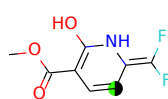
Low | 0.00   Low | 3.56   Medium | 10.39   Low | 0.00   Low | 6.29
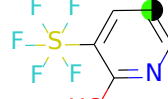
(S34.14) **398**   (S34.15) **427**
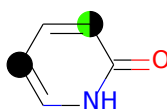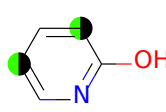
Medium | 0.00   Medium | 8.52

(S34.16) **226**

Pyridazines
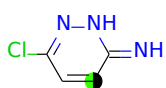
Medium | 0.00   Medium | 13.16   Medium | 0.00   Medium | 12.07   Medium | 0.00   Medium | 11.29

(S35.1) **32**   (S35.2) **33**   (S35.3) **34**

Pyrimidines

Medium | 0.00   Medium | 13.49

(S36.1) **25**

Medium | 0.00   Medium | 14.02

(S36.2) **26**

Medium | 0.00   Medium | 13.02

(S36.3) **27**

Medium | 0.00   Medium | 0.87

(S36.4) **28**

Medium | 0.00

(S36.5) **29**

Medium | 0.00   Medium | 13.51   High | 13.85

(S36.6) **85**

Medium | 0.00   Medium | 14.07

(S36.7) **86**

Medium | 0.00   High | 10.38   High | 13.39

(S36.8) **87**

Low | 0.00

(S36.9) **88**

Medium | 0.00   Medium | 12.98   Medium | 14.51

(S36.10) **89**

High | 0.00   High | 12.28

(S36.11) **90**

Medium | 0.00   Medium | 12.75   Medium | 13.24   Medium | 14.00

(S36.12) **91**

Medium | 0.00

(S36.13) **92**

Medium | 0.00   Medium | 12.75   High | 13.69   Medium | 13.98

(S36.14) **93**

Medium | 0.00   High | 13.34   Medium | 13.82   High | 14.72   Medium | 0.00   Medium | 13.15

(S36.15) **94**   (S36.16) **96**

Medium | 0.00   High | 10.18   Medium | 12.60   Medium | 13.18

Medium | 13.38   Medium | 14.83

(S36.17) **97**

Medium | 0.00   Medium | 8.31   Medium | 9.42   Medium | 12.76

Medium | 13.06   Medium | 13.60   Medium | 0.00   Medium | 13.35

(S36.18) **98**   (S36.19) **324**

Medium | 0.00   Medium | 8.55   Medium | 10.43   Medium | 14.40   Medium | 0.00

(S36.20) **390**   (S36.21) **593**

Medium | 0.00   Medium | 10.39   Medium | 12.77   Medium | 14.07

(S36.22) **597**

Pyrimidin-2(1H)-ones
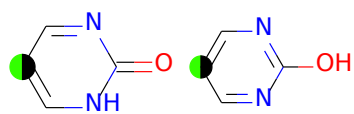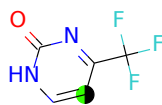
Low | 0.00   Low | 8.47   Low | 0.00   Low | 3.78   Low | 7.71
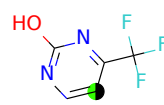
(S37.1) **99**   (S37.2) **100**

Pyrimidin-4(3H)-ones

Medium | 0.00   Medium | 3.05   Medium | 5.80   Medium | 6.70

(S38.1) **101**

Medium | 0.00   Medium | 3.54   Medium | 9.29   Medium | 0.00   Medium | 2.77   Medium | 8.46

(S38.2) **102**   (S38.3) **103**

Medium | 0.00   Medium | 5.45   Medium | 8.93   Low | 0.00   Medium | 2.93   Low | 8.26

(S38.4) **104**   (S38.5) **105**

Medium | 0.00   Medium | 3.03   Medium | 5.29   Medium | 6.01

(S38.6) **106**

Medium | 0.00   Medium | 6.02   Medium | 8.49   Medium | 0.00   Medium | 4.36   Medium | 7.98
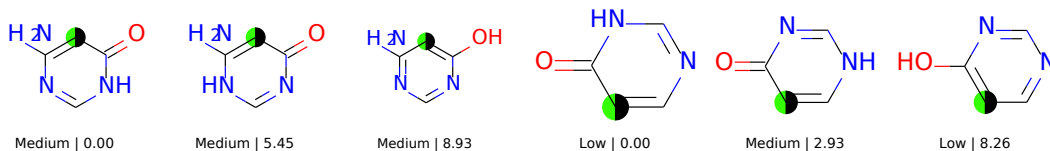
(S38.7) **594**   (S38.8) **595**
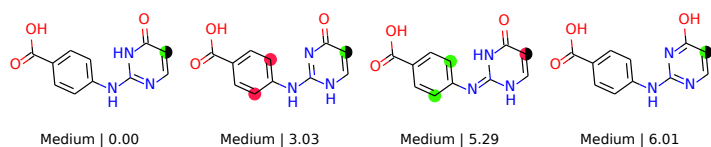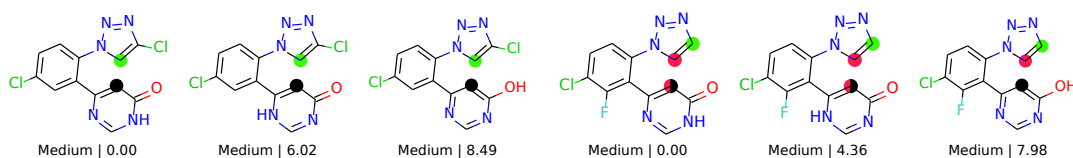
Medium | 0.00   Medium | 3.92   Medium | 7.80
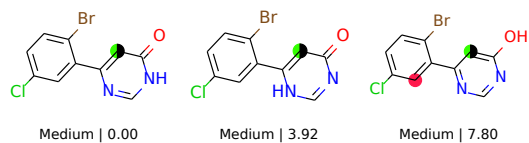
(S38.9) **596**

Pyrazines

Medium | 0.00

Medium | 0.00

Medium | 0.00

Medium | 13.60

Medium | 0.00

Medium | 14.72

(S39.1) **30** (S39.2) **31** (S39.3) **49** (S39.4) **107**

Medium | 0.00

Medium | 13.39

Medium | 0.00

Medium | 8.55

Medium | 0.00

Medium | 13.95

(S39.5) **108** (S39.6) **109** (S39.7) **131**

Medium | 0.00

Medium | 14.03

Medium | 0.00

Medium | 14.40

Medium | 0.00

Medium | 13.46

(S39.8) **132** (S39.9) **133** (S39.10) **134**

Medium | 0.00

Medium | 12.30

Medium | 0.00

Medium | 0.00

Medium | 0.00

Medium | 13.34

(S39.11) **135** (S39.12) **192** (S39.13) **193** (S39.14) **194**

Medium | 0.00

Medium | 12.37

Medium | 0.00

Medium | 13.57

Medium | 0.00

(S39.15) **196** (S39.16) **198** (S39.17) **199**

Medium | 0.00

Medium | 12.67

Medium | 0.00

Medium | 9.15

Medium | 0.00

Medium | 0.00

(S39.18) **200** (S39.19) **201** (S39.20) **202** (S39.21) **203**

Medium | 0.00   Medium | 9.06

(S39.22) **205**

Medium | 0.00   Medium | 12.60

(S39.23) **206**

Medium | 0.00   Medium | 12.72

(S39.24) **207**

Medium | 0.00   Medium | 0.00   Medium | 0.00   Medium | 9.92

(S39.25) **211**   (S39.26) **212**   (S39.27) **284**

Medium | 0.00   Medium | 9.57   Medium | 12.68   Medium | 0.00   Medium | 13.84

(S39.28) **285**   (S39.29) **286**
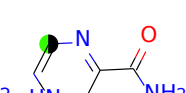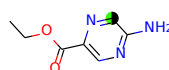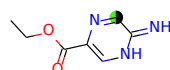
Medium | 0.00   Medium | 11.73   Medium | 0.00   Medium | 12.90   Medium | 0.00   Medium | 13.33
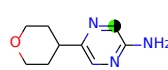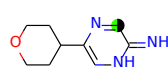
(S39.30) **289**   (S39.31) **290**   (S39.32) **291**
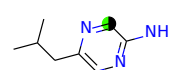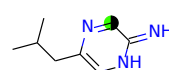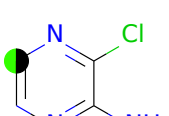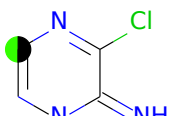
Medium | 0.00   Medium | 12.65   Medium | 0.00   Medium | 0.00   Medium | 12.72   Medium | 0.00
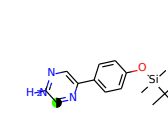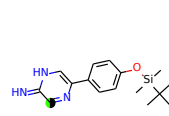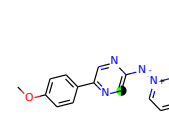
(S39.33) **292**   (S39.34) **293**   (S39.35) **195**   (S39.36) **209**
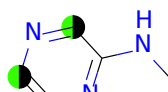
Medium | 0.00   Medium | 10.41

(S39.37) **283**

Pyrazinones

Medium | 0.00   Medium | 7.09   Medium | 0.00   Medium | 2.63   Low | 8.10

(S40.1) **190**   (S40.2) **197**

Medium | 0.00   Medium | 1.57   Medium | 9.71   Medium | 12.14

Medium | 14.12

(S40.3) **204**
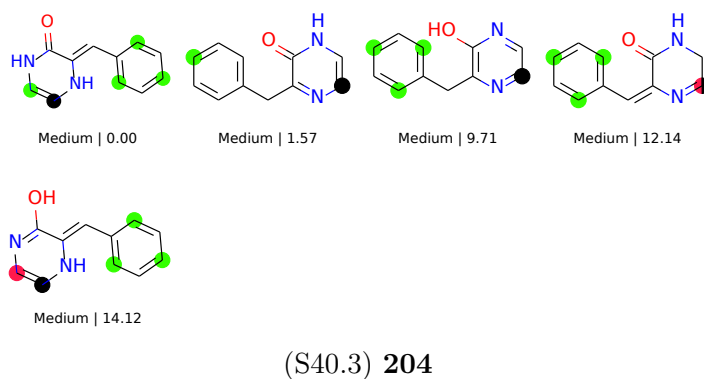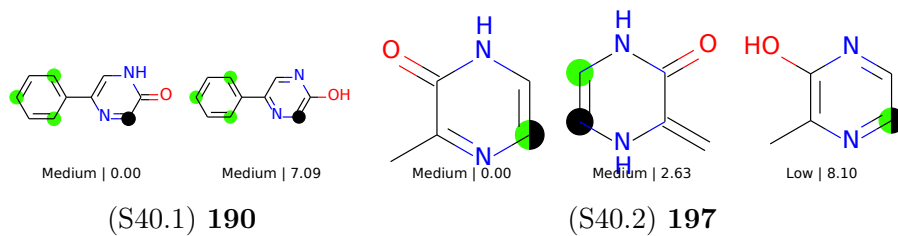
4H-Furo[3,2-b]pyrroles

Medium | 0.00

(S41.1) **472**

4H-Thieno[3,2-b]pyrroles

Medium | 0.00

(S42.1) **471**
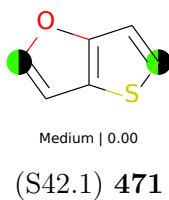
Imidazo[2,1-b][1,3,4]thiadiazoles
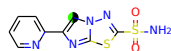
Medium | 0.00

(S43.1) **418**

Indoles
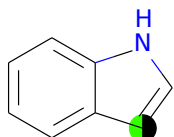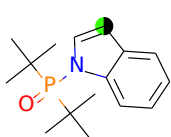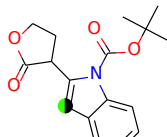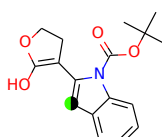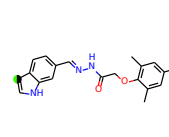
| Medium \| 0.00 | Medium \| 0.00 | Medium \| 0.00 | High \| 13.69 | Medium \| 0.00 | Medium \| 13.94 |
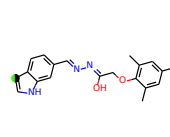
(S44.1) **35**   (S44.2) **399**   (S44.3) **407**   (S44.4) **412**
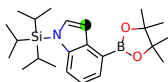
Medium | 0.00   Medium | 0.00
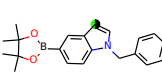
(S44.5) **425**   (S44.6) **426**

Benzofurans
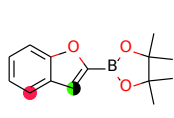
Medium | 0.00   Medium | 0.00   Medium | 0.00   Medium | 0.00   Medium | 0.00   Medium | 0.00
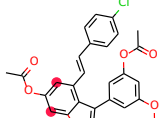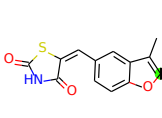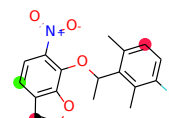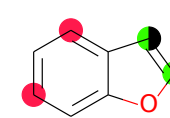
(S45.1) **424**   (S45.2) **430**   (S45.3) **431**   (S45.4) **461**   (S45.5) **468**   (S45.6) **470**
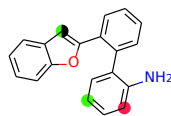
Medium | 0.00

(S45.7) **433**

Benzo[b]thiophenes
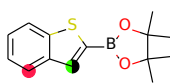
Medium | 0.00

(S46.1) **423**

Medium | 0.00

(S46.2) **429**

Indazoles

Medium | 0.00   Medium | 3.49   Medium | 0.00   Medium | 3.31   High | 13.74

(S47.1) **353**   (S47.2) **354**

Medium | 0.00   Medium | 3.33   Medium | 0.00   Medium | 3.31   Medium | 0.00   Medium | 3.12
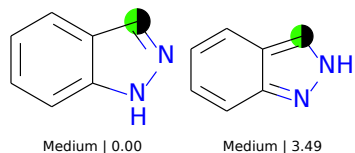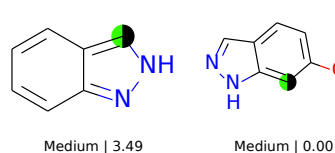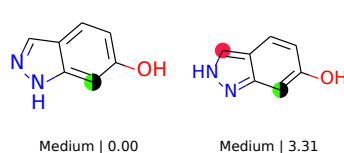
(S47.3) **355**   (S47.4) **356**   (S47.5) **357**

Medium | 0.00   Medium | 3.15   Medium | 0.00   Medium | 2.92   Medium | 0.00   Medium | 1.48
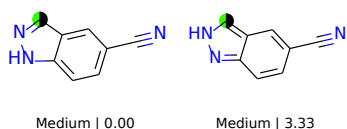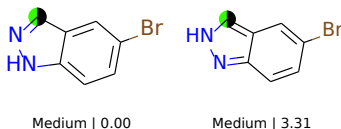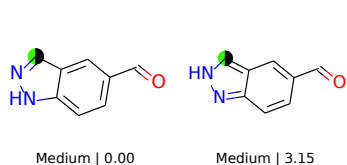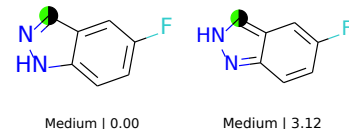
(S47.6) **358**   (S47.7) **359**   (S47.8) **360**

Medium | 0.00   Medium | 3.39   Medium | 0.00   Medium | 3.06   Medium | 0.00   Medium | 2.64
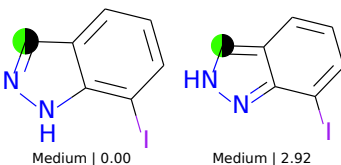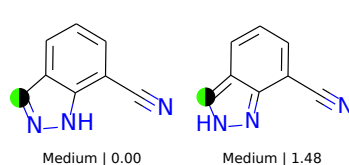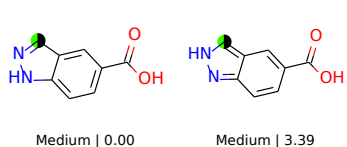
(S47.9) **361**   (S47.10) **362**   (S47.11) **363**

Medium | 0.00   Medium | 0.00   Medium | 2.70   Medium | 0.00   Medium | 2.82
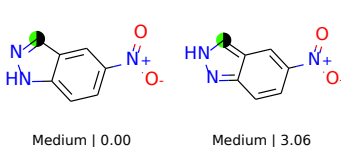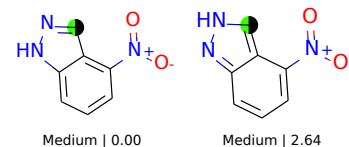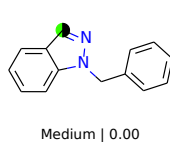
(S47.12) **364**   (S47.13) **365**   (S47.14) **366**

Medium | 0.00   High | 14.87   Medium | 0.00   Medium | 3.55   Medium | 0.00   Medium | 3.69
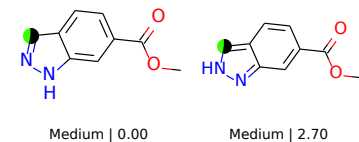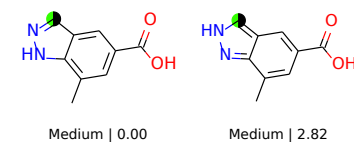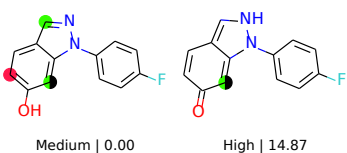
(S47.15) **370**   (S47.16) **371**   (S47.17) **372**

Medium | 0.00   High | 14.00   Medium | 0.00   Medium | 4.13   Medium | 0.00   Medium | 2.87
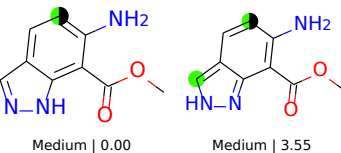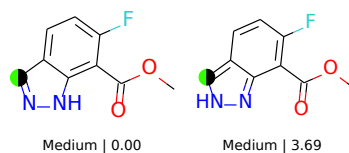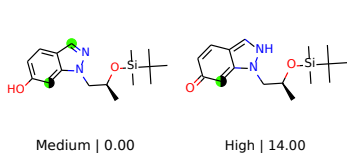
(S47.18) **374**   (S47.19) **375**   (S47.20) **383**

Medium | 0.00

Medium | 3.22

Medium | 0.00

Medium | 6.46

Medium | 0.00

(S47.21) **367**

(S47.22) **368**

(S47.23) **369**

Medium | 0.00

Medium | 3.36

(S47.24) **373**

1H-Benzo[d]imidazoles

Medium | 0.00

Medium | 0.07

Medium | 7.61

Medium | 0.00

Medium | 0.85

(S48.1) **377**

(S48.2) **378**

Medium | 0.00

Medium | 1.24

Medium | 0.00

Medium | 4.52

Medium | 0.00

Medium | 2.51

(S48.3) **381**

(S48.4) **382**

(S48.5) **385**

Medium | 0.00

Medium | 1.70

Medium | 0.00

Medium | 4.91

(S48.6) **387**

(S48.7) **388**

7-Azaindoles

Medium | 0.00   Medium | 5.12   Medium | 0.00   High | 10.86   Medium | 0.00   Medium | 0.00
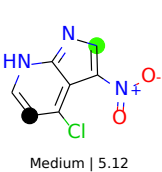
(S49.1) **334**   (S49.2) **335**   (S49.3) **336**   (S49.4) **337**
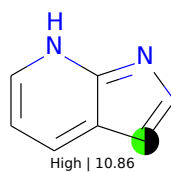
Medium | 0.00   Medium | 0.00   High | 12.10   Medium | 0.00   High | 11.77
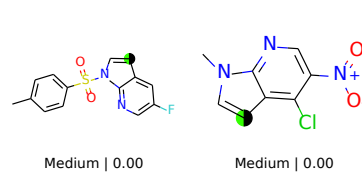
(S49.5) **338**   (S49.6) **339**   (S49.7) **340**

Medium | 0.00   High | 11.35   Medium | 0.00   High | 8.31

(S49.8) **341**   (S49.9) **342**

Medium | 0.00   Medium | 4.27   Medium | 9.98   High | 11.50   Medium | 0.00   Medium | 0.00
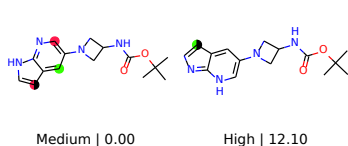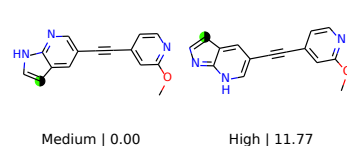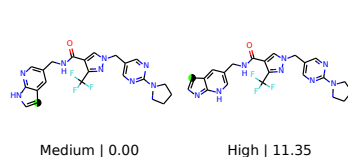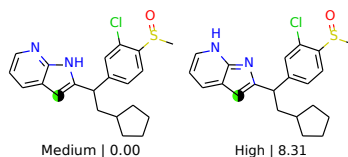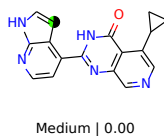
(S49.10) **343**   (S49.11) **344**   (S49.12) **345**

High | 0.00   High | 7.00   Medium | 0.00   Medium | 0.00   Medium | 0.00
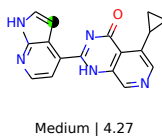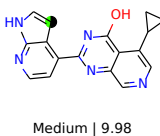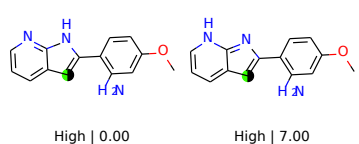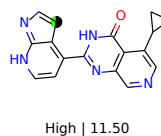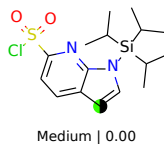
(S49.13) **346**   (S49.14) **347**   (S49.15) **348**   (S49.16) **349**

Medium | 0.00   High | 9.14   Medium | 12.44   Medium | 14.51

(S49.17) **351**

6-Azaindoles

Medium | 0.00   Medium | 8.66

(S50.1) **316**

Medium | 0.00   High | 11.40

(S50.2) **322**

Medium | 0.00   High | 10.56

(S50.3) **323**

Medium | 0.00   High | 10.15

(S50.4) **325**

Medium | 0.00   High | 9.11

(S50.5) **326**

Medium | 0.00   High | 14.28

(S50.6) **327**

Medium | 0.00

(S50.7) **328**

Medium | 0.00

(S50.8) **329**
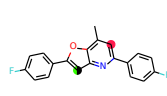
Medium | 0.00

(S50.9) **330**

Medium | 0.00

(S50.10) **331**

Medium | 0.00

(S50.11) **332**

Medium | 0.00   High | 12.37

(S50.12) **333**

5-Azaindoles

Medium | 0.00   High | 12.47

(S51.1) **317**

Medium | 0.00   High | 13.86

(S51.2) **318**

High | 0.00   High | 12.40

(S51.3) **319**

Medium | 0.00   High | 13.60

(S51.4) **320**

Medium | 0.00   Medium | 1.76   High | 11.41

(S51.5) **321**

59

4-Azaindoles

Medium | 0.00   High | 4.01

(S52.1) **300**

Medium | 0.00   Medium | 7.88   Medium | 11.71

(S52.2) **301**

Medium | 0.00   Medium | 7.68   Medium | 8.11   Medium | 14.20

(S52.3) **302**

Medium | 0.00   Medium | 7.97   Medium | 8.53   Medium | 14.10   Medium | 0.00

(S52.4) **303**   (S52.5) **304**

Medium | 0.00   High | 9.08   Medium | 0.00   High | 9.43   Medium | 0.00   High | 9.53

(S52.6) **305**   (S52.7) **306**   (S52.8) **307**

Medium | 0.00   Medium | 0.00   Medium | 0.00   High | 11.01   Medium | 0.00
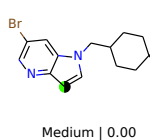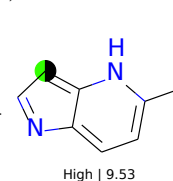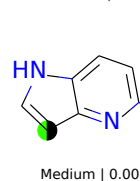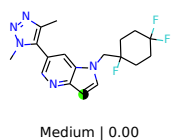
(S52.9) **308**   (S52.10) **309**   (S52.11) **310**   (S52.12) **311**

Medium | 0.00   High | 10.10   Medium | 0.00   High | 11.91   Medium | 12.47   Medium | 14.63
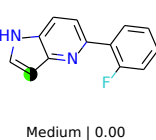
(S52.13) **312**   (S52.14) **313**
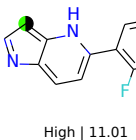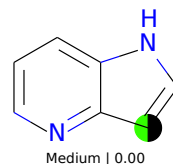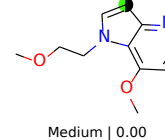
Medium | 0.00   Medium | 0.00   High | 11.69
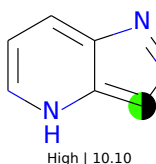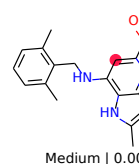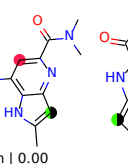
(S52.15) **314**   (S52.16) **315**

Imidazo[1,2-a]pyridines

High | 0.00

(S53.1) **170**

Medium | 0.00

(S53.2) **172**

Medium | 0.00

Medium | 11.13

Medium | 11.37

High | 11.45

(S53.3) **174**

Medium | 0.00

Medium | 12.12

(S53.4) **176**

Medium | 0.00

Medium | 11.95

(S53.5) **177**

Medium | 0.00

(S53.6) **178**

Medium | 0.00

(S53.7) **179**

Medium | 0.00

(S53.8) **180**

Medium | 0.00

(S53.9) **182**

Medium | 0.00

(S53.10) **184**

Medium | 0.00

(S53.11) **187**

Medium | 0.00

(S53.12) **188**

High | 0.00

(S53.13) **189**

Medium | 0.00

(S53.14) **408**

Furo[2,3-b]pyridines

Medium | 0.00

(S54.1) **455**

Medium | 0.00

(S54.2) **462**
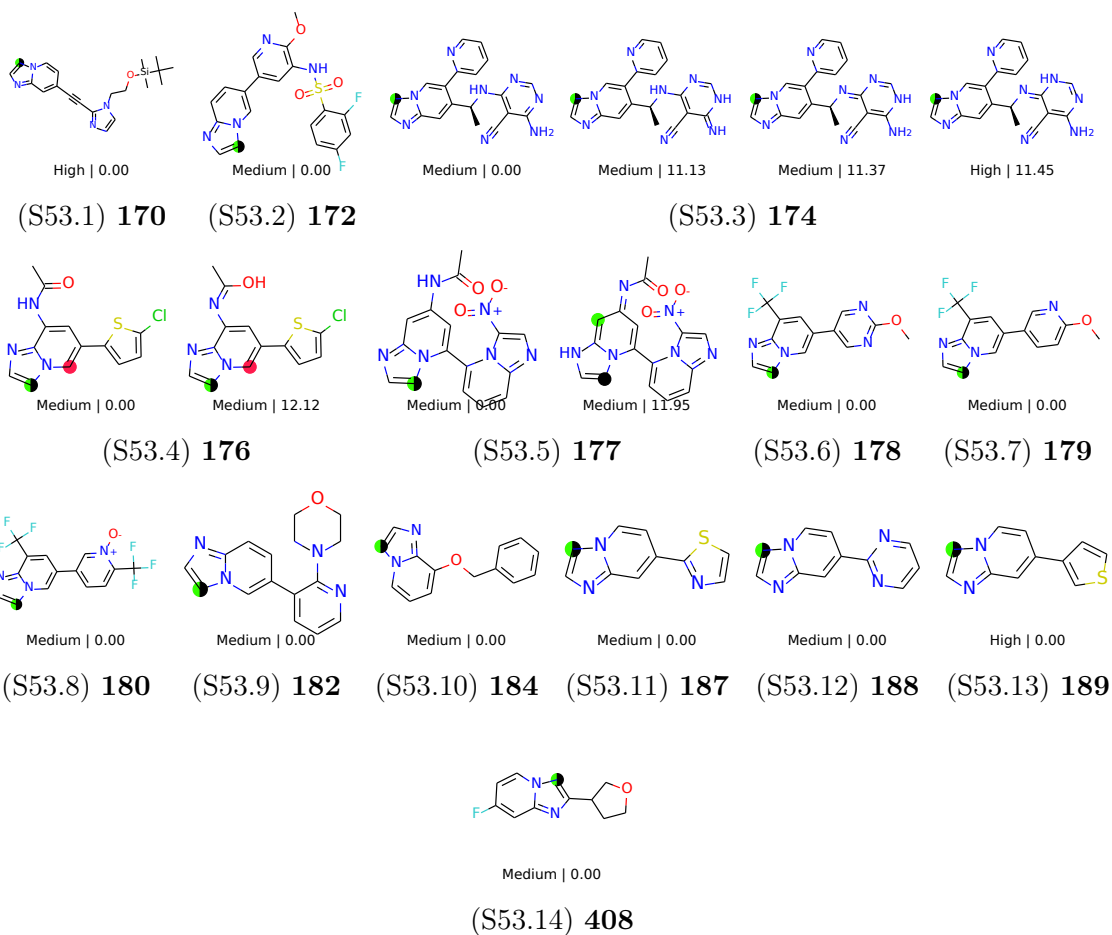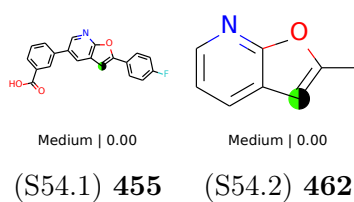
Furo[2,3-c]pyridines

Medium | 0.00   Medium | 10.08
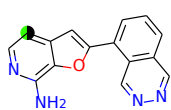
(S55.1) **268**
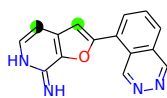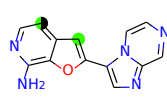
Medium | 0.00   Medium | 10.10

(S55.2) **269**

Medium | 0.00   Medium | 10.53
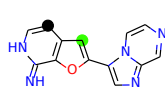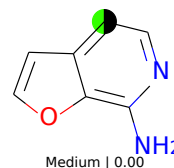
(S55.3) **270**
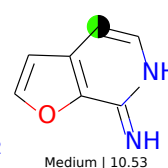
Medium | 0.00   Medium | 9.42
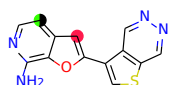
(S55.4) **271**
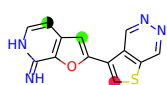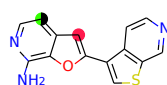
Medium | 0.00   Medium | 9.81

(S55.5) **272**
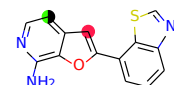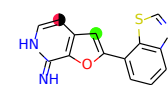
Medium | 0.00   Medium | 9.13

(S55.6) **273**
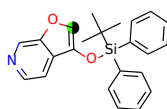
Medium | 0.00

(S55.7) **473**
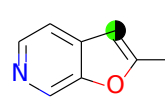
Medium | 0.00

(S55.8) **480**

Furo[3,2-c]pyridines
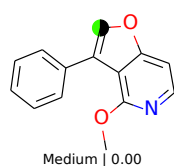
Medium | 0.00

(S56.1) **459**
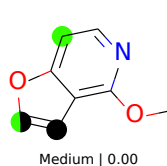
Medium | 0.00

(S56.2) **460**
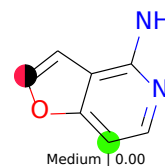
Medium | 0.00   Medium | 10.41

(S56.3) **482**
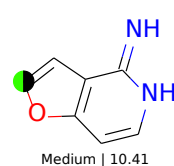
Furo[3,2-b]pyridines
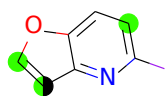
Medium | 0.00

(S57.1) **439**

Medium | 0.00

(S57.2) **440**
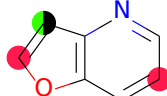
Medium | 0.00

(S57.3) **447**
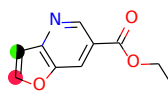
Low | 0.00

(S57.4) **448**
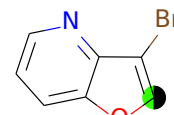
Medium | 0.00

(S57.5) **481**

[1,2,4]Triazolo[4,3-a]pyridines

Medium | 0.00   Medium | 0.00

(S58.1) **143**   (S58.2) **150**
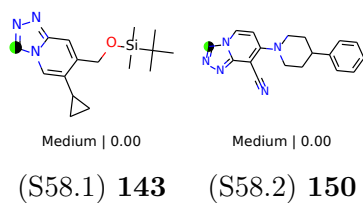
1H-Pyrazolo[3,4-b]pyridines

Medium | 0.00   Medium | 1.60   Medium | 14.37

(S59.1) **394**
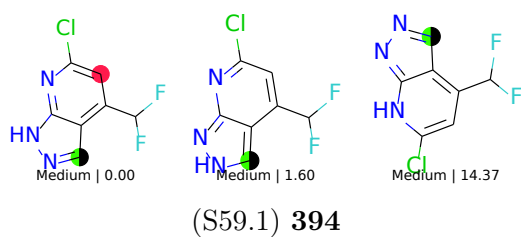
1H-Imidazo[4,5-c]pyridines

Medium | 0.00

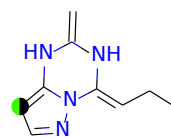(S60.1) **265**

Imidazo[1,2-a]pyrazines

Medium | 0.00   Medium | 0.00   Medium | 0.00   Medium | 0.00   Medium | 0.00   Medium | 0.00

(S61.1) **38**   (S61.2) **39**   (S61.3) **40**   (S61.4) **111**   (S61.5) **112**   (S61.6) **186**
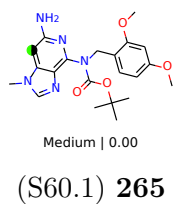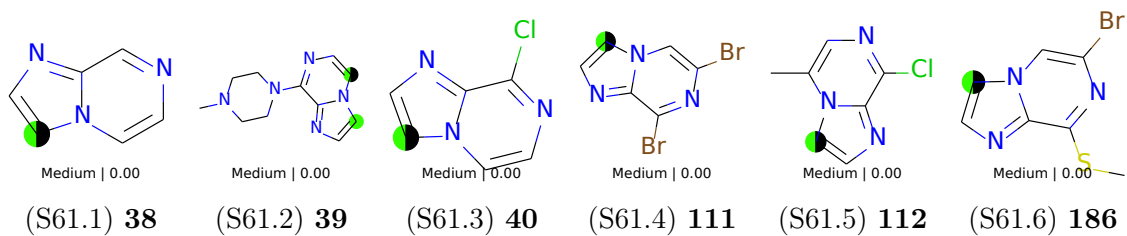
Imidazo[1,2-c]pyrimidines

High | 0.00
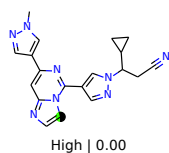
(S62.1) **175**

Imidazo[1,2-b]pyridazines

High | 0.00
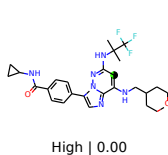
(S63.1) **280**
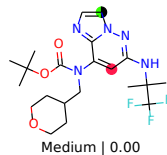
Medium | 0.00

(S63.2) **281**

7H-Pyrrolo[2,3-d]pyrimidines

Medium | 0.00

(S64.1) **275**

Medium | 0.00
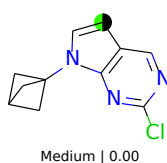
(S64.2) **403**
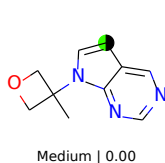
Medium | 0.00
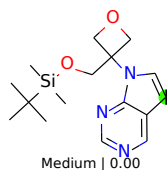
(S64.3) **404**
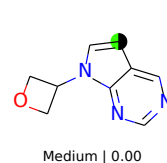
Medium | 0.00

(S64.4) **405**

Pyrazolo[1,5-c]pyrimidines

Medium | 0.00

(S65.1) **36**

High | 7.42

Medium | 0.00

(S65.2) **115**

Medium | 7.14

Imidazo[1,5-a]pyrazines

Medium | 0.00   High | 11.67   Medium | 0.00   Medium | 2.80

(S66.1) **279**   (S66.2) **406**
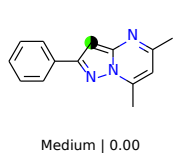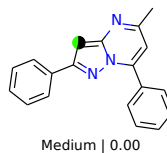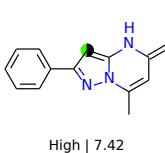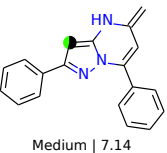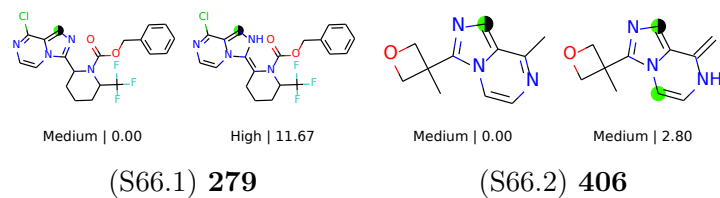
Pyrrolo[2,1-f][1,2,4]triazines

Medium | 0.00   Medium | 7.18

(S67.1) **411**

Oxazolo[4,5-b]pyridines

Medium | 0.00

(S68.1) **264**

Furo[2,3-d]pyrimidines

Medium | 0.00   Medium | 0.00   Medium | 0.00   Medium | 0.00   Medium | 10.57

(S69.1) **457**   (S69.2) **463**   (S69.3) **464**   (S69.4) **467**
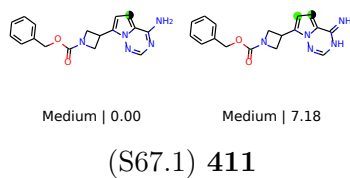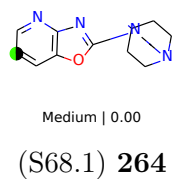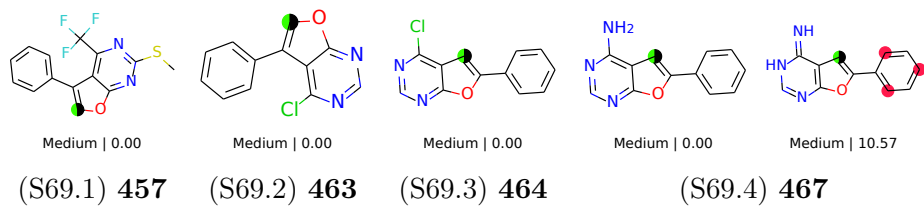
Furo[2,3-b]pyrazines
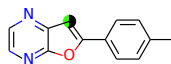
Medium | 0.00

(S70.1) **487**

[1,2,4]Triazolo[4,3-b]pyridazines
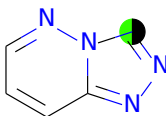
Medium | 0.00

(S71.1) **41**
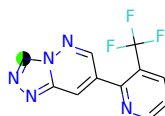
Medium | 0.00

(S71.2) **116**

Pyrazolo[1,5-a][1,3,5]triazines
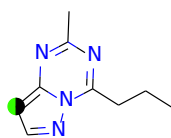
Medium | 0.00
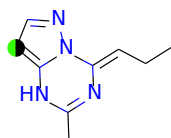
Medium | 0.29
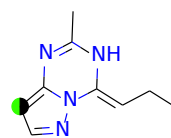
Medium | 0.65

Medium | 5.96
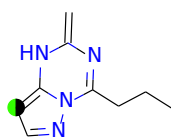
Medium | 6.11
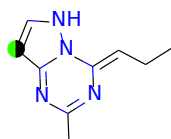
Medium | 14.83
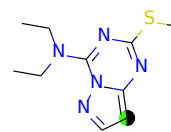
(S72.1) **113**

Medium | 0.00

(S72.2) **114**
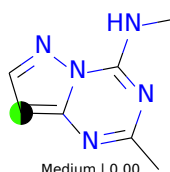
Medium | 0.00
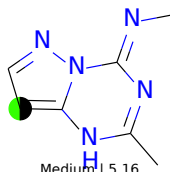
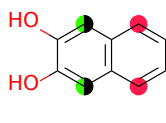Medium | 5.16
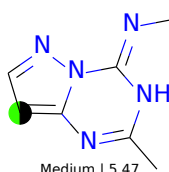
Medium | 5.47
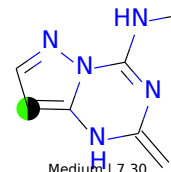
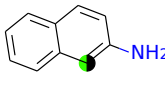Medium | 7.30
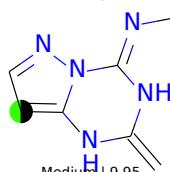
Medium | 9.95

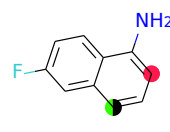(S72.3) **117**

Naphthalenes
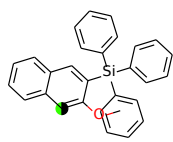
Medium | 0.00

(S73.1) **510**

Medium | 0.00

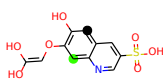(S73.2) **512**

Medium | 0.00

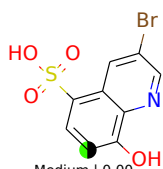(S73.3) **538**

Medium | 0.00

(S73.4) **544**
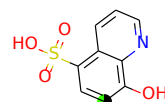
Quinolines
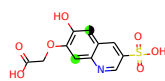
Medium | 0.00

Medium | 15.00
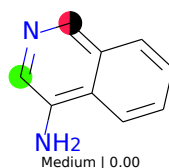
(S74.1) **415**

Medium | 0.00

(S74.2) **416**
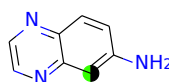
Medium | 0.00

(S74.3) **417**

Isoquinolines

Medium | 0.00

(S75.1) **266**

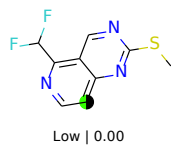# Quinoxalines

Medium | 0.00

(S76.1) **548**

1,5-Naphthyridines

Medium | 0.00   Medium | 6.81
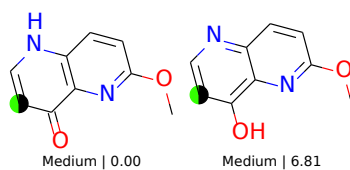
(S77.1) **260**

Pyrido[4,3-d]pyrimidines

Low | 0.00

(S78.1) **396**

Pyrido[3,4-b]pyrazines

Low | 0.00
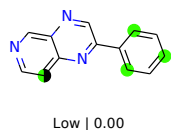
(S79.1) **262**

Furo[3,2-c]pyridin-4(5H)-ones

Medium | 0.00   Medium | 9.43
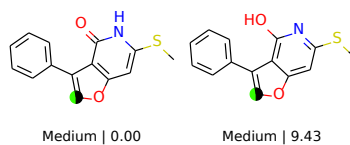
(S80.1) **456**

Furo[2,3-d]pyrimidin-4(3H)-ones

Medium | 0.00   Medium | 0.00

(S81.1) **445**   (S81.2) **458**
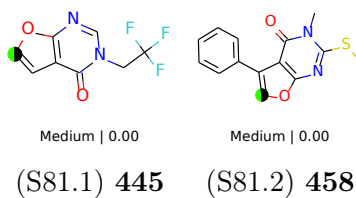
Imidazo[1,2-a]pyrazin-8(7H)-ones

Medium | 0.00   Medium | 11.76

(S82.1) **181**
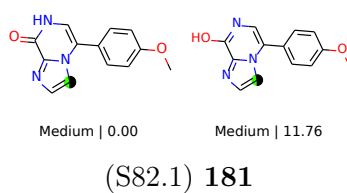
Thiazolo[5,4-b]pyridin-5(4H)-ones

Medium | 0.00   Medium | 10.57

(S83.1) **585**
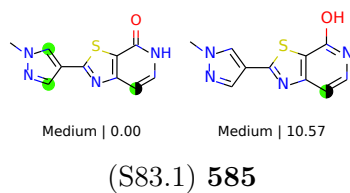
3,4-Dihydro-5H-[1,2,3]triazolo[4,5-b]pyridin-5-ones

Medium | 0.00   Medium | 4.14

(S84.1) **267**
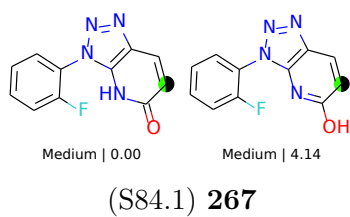
1,3-Dihydro-2H-benzo[d]imidazol-2-ones

Medium | 0.00

(S85.1) **376**

Medium | 0.00

(S85.2) **386**

Medium | 0.00

(S85.3) **389**

Medium | 0.00

(S85.4) **384**
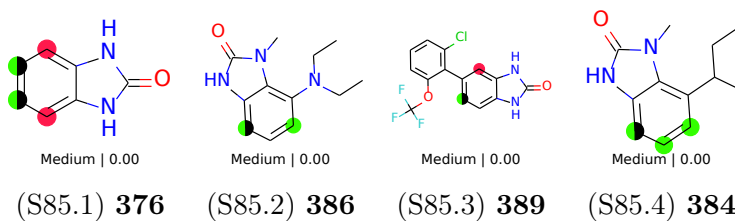
Oxazolo[4,5-b]pyridin-2(3H)-ones

Low | 0.00

Low | 4.93

(S86.1) **261**

Quinazolin-4(3H)-ones

Medium | 0.00

(S87.1) **580**

Medium | 0.00

(S87.2) **581**
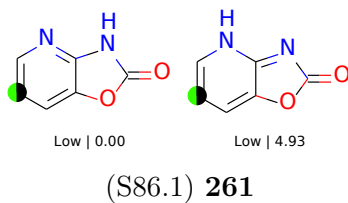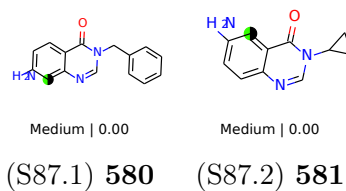
1,7-Naphthyridin-8(7H)-ones

Medium | 0.00

Medium | 8.30

Medium | 11.47

(S88.1) **263**
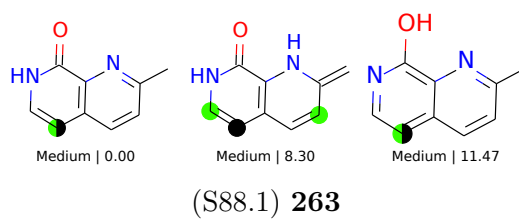
Pyrido[2,3-d]pyrimidin-7(8H)-ones

Medium | 0.00   Medium | 4.30   Medium | 14.02   Medium | 14.14
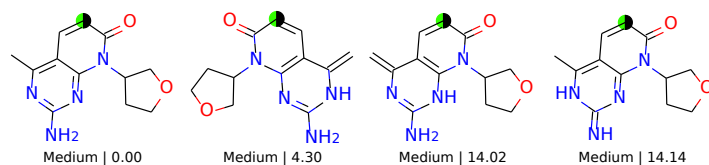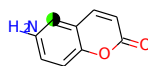
(S89.1) **409**

2H-Chromen-2-ones

Medium | 0.00

(S90.1) **557**

Graphical TOC Entry