



LJMU Research Online

Kramer, C, Ting, A, Zheng, H, Hert, J, Schindler, T, Stahl, M, Robb, G, Crawford, JJ, Blaney, J, Montague, S, Leach, AG, Dossetter, AG and Griffen, EJ

Learning Medicinal Chemistry Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) Rules from Cross-Company Matched Molecular Pairs Analysis (MMPA).

<http://researchonline.ljmu.ac.uk/id/eprint/7283/>

Article

Citation (please note it is advisable to refer to the publisher's version if you intend to cite from this work)

Kramer, C, Ting, A, Zheng, H, Hert, J, Schindler, T, Stahl, M, Robb, G, Crawford, JJ, Blaney, J, Montague, S, Leach, AG, Dossetter, AG and Griffen, EJ (2017) Learning Medicinal Chemistry Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) Rules from Cross-Company

LJMU has developed **LJMU Research Online** for users to access the research output of the University more effectively. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Users may download and/or print one copy of any article(s) in LJMU Research Online to facilitate their private study or for non-commercial research. You may not engage in further distribution of the material or use it for any profit-making activities or any commercial gain.

The version presented here may differ from the published version or from the version of the record. Please see the repository URL above for details on accessing the published version and note that access may require a subscription.

For more information please contact researchonline@ljmu.ac.uk

<http://researchonline.ljmu.ac.uk/>

This document is confidential and is proprietary to the American Chemical Society and its authors. Do not copy or disclose without written permission. If you have received this item in error, notify the sender and delete all copies.

Learning Medicinal Chemistry Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) rules from Cross-company Matched Molecular Pairs Analysis (MMPA)

Journal:	<i>Journal of Medicinal Chemistry</i>
Manuscript ID	jm-2017-00935p.R2
Manuscript Type:	Perspective
Date Submitted by the Author:	12-Sep-2017
Complete List of Authors:	Kramer, Christian; F. Hoffmann-La Roche AG, Research and Development Division Ting, Attila; AstraZeneca plc Zheng, Hao; Genentech Inc Hert, Jérôme; F Hoffmann-La Roche AG Research and Development Division, Roche Innovation Center Basel Schindler, Torsten; F Hoffmann-La Roche AG Research and Development Division, Roche Innovation Center Basel Stahl, Martin; F Hoffmann-La Roche AG Research and Development Division, Roche Innovation Center Basel Robb, Graeme; AstraZeneca plc Crawford, James; Genentech Inc Blaney, Jeff; Genentech Inc Montague, Shane; MedChemica Limited Leach, Andrew; MedChemica Limited Dossetter, Alexander; MedChemica Limited Griffen, Edward; MedChemica Limited

SCHOLARONE™
Manuscripts

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Learning Medicinal Chemistry

Absorption, Distribution,

Metabolism, Excretion, and

Toxicity (ADMET) rules from

Cross-company Matched

Molecular Pairs Analysis (MMPA)

Christian Kramer^{*,§}, Attila Ting^{*,*}, Hao Zheng^{*,‡}, Jerome Hert[§], Torsten Schindler[§],
Martin Stahl[§], Graeme Robb^{*}, James J. Crawford[‡], Jeff Blaney[‡], Shane Montague[†], Andrew
G. Leach[†], Al. G. Dossetter[†], Ed J. Griffen^{*,†}

[§]Roche Pharma Research and Early Development, Roche Innovation Center Basel,
Switzerland, [‡]Genentech Inc, 1 DNA Way, South San Francisco, CA 94080, ^{*}AstraZeneca
plc, Milton Rd, Milton, Cambridge, CB4 0FZ, [†]MedChemica Ltd, Biohub Alderley Park,
Macclesfield. Cheshire SK10 4TG.

Abstract

1
2
3 The first large scale analysis of in vitro Absorption, Distribution, Metabolism, Excretion,
4 and Toxicity (ADMET) data shared across multiple major Pharma has been performed. Using
5 advanced matched molecular pair analysis (MMPA), we combined data from three
6 pharmaceutical companies and generated ADMET rules, avoiding the need to disclose the
7 full chemical structures. On top of the very large exchange of knowledge, all companies
8 involved synergistically gained approximately 20% more rules from the shared
9 transformations. There is good quantitative agreement between the rules based on shared data
10 compared to both individual companies' rules and rules published in the literature. Known
11 correlations between logD, solubility, in vitro clearance and plasma protein binding also hold
12 in transformation space, but there are also interesting exceptions. Data pools such as this
13 allow focusing on particular functional groups and characterizing their ADMET profile.
14 Finally the role of a corpus of robustly tested medicinal chemistry knowledge in the training
15 of medicinal chemistry is discussed.
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31

32 ***Keywords***

33
34
35
36
37 ADMET, MMPA, matched molecular pair analysis, data sharing, data mining, big data,
38 unsupervised learning, precompetitive research, medicinal chemistry knowledge.
39
40
41
42

43 ***Introduction***

44
45
46 Medicinal chemistry is a discipline at the borderline of science and technology that
47 depends on experience, intuition and knowledge of rules that govern the space of medicinal
48 chemistry practice. As such, successful medicinal chemistry optimization has routinely drawn
49 on knowledge gained from past experimental observations. Thus it is fair to assume that the
50 more data are available (everything else remaining equal), the more knowledge can be
51 gained.
52
53
54
55
56
57
58
59
60

1
2
3 While more data is desirable, the strategic necessities in a patent-based competitive
4 environment allow the public sharing of only a small proportion all of the information
5 generated inside pharmaceutical companies. In particular, the composition of chemical matter
6 with its associated biological properties is a cornerstone of pharmaceutical patent filing.
7 Therefore parties interested in protecting their research must exercise great care in the
8 chemical structures they disclose either in a patent or in academic publications not to weaken
9 any patent position. It is therefore unsurprising that only a small proportion of chemical
10 structures synthesized and their assay data tested are ever made public. As medicinal
11 chemistry knowledge is the understanding of the relationships between chemical structure
12 and biological properties, this creates a conflict between gaining knowledge and maintaining
13 intellectual property.
14
15
16
17
18
19
20
21
22
23
24
25
26

27
28 In order to increase our knowledge space, we set out on an experiment to combine
29 absorption, distribution, metabolism, and toxicity (ADMET) data from three different large
30 pharmaceutical companies (AstraZeneca, Genentech, and Roche; Genentech research and
31 Roche research operate independently) and mine the pooled data for relevant medicinal
32 chemistry knowledge. MedChemica was used as the intermediary to combine rules. In
33 particular, we were interested in whether we could merge ADMET knowledge between the
34 companies, and whether we could synergistically increase our medicinal chemistry
35 knowledge based upon the joint information pool.
36
37
38
39
40
41
42
43
44
45
46

47 A solution to sharing large amounts of medicinal chemistry data between companies in
48 the patent-based competitive environment is to use matched molecular pair analysis
49 (MMPA),¹⁻⁷ since the original structures of the compounds cannot be calculated back if only
50 the part of the structure which is involved in the transformation is shared.⁸ MMPA extracts
51 transformations that link a pair of compounds. For all pairs that are linked by the same
52
53
54
55
56
57
58
59
60

1
2
3 transformation, MMPA calculates aggregate statistics of the property differences to derive a
4 rule for the given transformation. In this work we use the term “rule” to indicate a “virtual”
5 chemical transformation leading to a change in a physical chemistry or biological endpoint
6 where there are enough examples to support an assertion as to how this change influences
7 certain given molecular properties. Very rarely, there are transformations where all known
8 pairs change property in the same direction, in particular if there are many pairs.
9 Nevertheless, even in the presence of exceptions, knowing the most common effects is useful
10 to guide whether or not a compound shall be made. The rules therefore capture variability,
11 which is a crucial element in medicinal chemistry decision making. Obviously more than one
12 pair is needed for a rule to be more than an anecdote; the statistical approach used here
13 suggests an absolute minimum of six pairs should be considered in order to form a rule. One
14 of the most important attributes of MMPA is its closeness to medicinal chemistry reasoning,
15 where talk of “adding a fluorine to block metabolism” or “adding a solubilizing group” are
16 common descriptions of underlying thinking. Occam’s razor suggests that the simplest cause
17 for a change in a property is the single change in structure. Neighboring groups within the
18 molecules involved may also play a causal role, hence, the local chemical environment is
19 incorporated into the encoding of the structural change and should allow these effects to be
20 partitioned and examined separately.⁹ The only remaining explanations are long range effects
21 or non-additivity of chemical changes. Kramer et al showed that SAR non-additivity may be
22 masked due to experimental uncertainty in many cases, and in other cases depends on the
23 target and binding mode.¹⁰

24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51 There have been significant studies, already published, describing the application of
52 MMPA to ADMET datasets from individual pharmaceutical companies.^{2,11,12} Most of these
53 have been supervised analyses, where a particular question is asked such as “what effect does
54 hydrogen replacement on an aryl ring have on ADMET properties”^{13,14} or “what are the
55
56
57
58
59
60

1
2
3 effects of substituting different heterocycles for phenyls".¹² In contrast, here we attempt to
4
5 characterize the entire ADMET rule space.
6
7

8
9 Having developed the technology to merge data between companies (details can be found
10 in supporting information), the next question becomes: "How many extra rules do we gain
11 and is there any degradation in knowledge by merging data between companies?" It is
12 relatively obvious that by aggregating pairs from different companies the statistics for some
13 rarer transformations will become significant because there are more pairs. However, this
14 gain may be obfuscated by noise introduced due to mixing data from assays with the same
15 endpoint but run under different conditions.¹⁵ A major concern is assay comparability:
16 although two companies may formally measure the same ADMET attribute, if the assays
17 used are different the outcomes may be different. The advantage of MMPA here is that two
18 compounds are only compared if they have been measured by the same operator using the
19 same method and endpoint. The difference in an attribute between two compounds is then
20 calculated. If the assays are merely systematically displaced against other, then the systematic
21 displacement is automatically factored out by using differences rather than absolute values. It
22 is then possible to test if pairs can be aggregated across different assays and rules identified.
23 To validate this, we (a) compare the rules we found by aggregating pair sets across
24 companies to rules obtained from one company only and (b) to rules previously published.
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44

45 High-quality analysis from MMPA requires large amounts of data, and we will show that
46 more data is clearly better with brief examples of analyses that can be made using such a
47 massive dataset. We will illustrate the types of analyses that can be done having a rule pool as
48 large as ours. We initially present general analyses of correlations of solubility, clearance,
49 and PPB rules with logD (pH 7.4) rules. As rules representing general knowledge are always
50 accompanied by exceptions, we highlight a few less-intuitive medicinal chemistry rules that
51
52
53
54
55
56
57
58
59
60

1
2
3 counter the general trends. Finally, we give a taste of how such a huge set of rules can be
4
5 used to characterize individual functional groups, a statistical type of analysis which only
6
7 becomes possible once large MMP datasets are available. It is not our intent to hypothesize
8
9 the cause of a particular quoted rule being significant as a thorough exploration of the rules
10
11 has been beyond the scope of this work, and value can be obtained from their knowledge
12
13 without complete clarity as to their causation.
14
15
16
17

18 ***Methods***

20 ***Data extraction***

21
22
23 Data needs to be transformed into a normally distributed variable before MMPA *i.e.*
24
25 pIC_{50} , \log_{10} (molar solubility), \log_{10} (intrinsic Clearance, (CL_{int})) for in vitro clearance assays,
26
27 \log (free/bound) for protein binding etc. Out of range flags were attached where appropriate.
28
29 If there were multiple measurements in the same assay for the same compound then the data
30
31 were aggregated using the following scheme:
32
33
34
35
36

- 37 1. Where there are no qualifiers / out of range flags:
 - 38 · aggregate as median logged value – as this is more representative of true value
 - 39
 - 40
 - 41
 - 42
 - 43
 - 44
 - 45
 - 46
 - 47
 - 48
 - 49
 - 50
 - 51
- 52 2. Where there are some qualified data present:
 - 53 · take the median of the data excluding qualified data:
 - 54
 - 55
 - 56
 - 57

58 For example: 3, 6, >8, >10, <3, 5, 6: exclude >8, >10, <3, the
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

- if there is a mixture of > and < then discard the data
- if the qualifiers are all < then use the minimum value
- if the qualifiers are all > then use the maximum value.

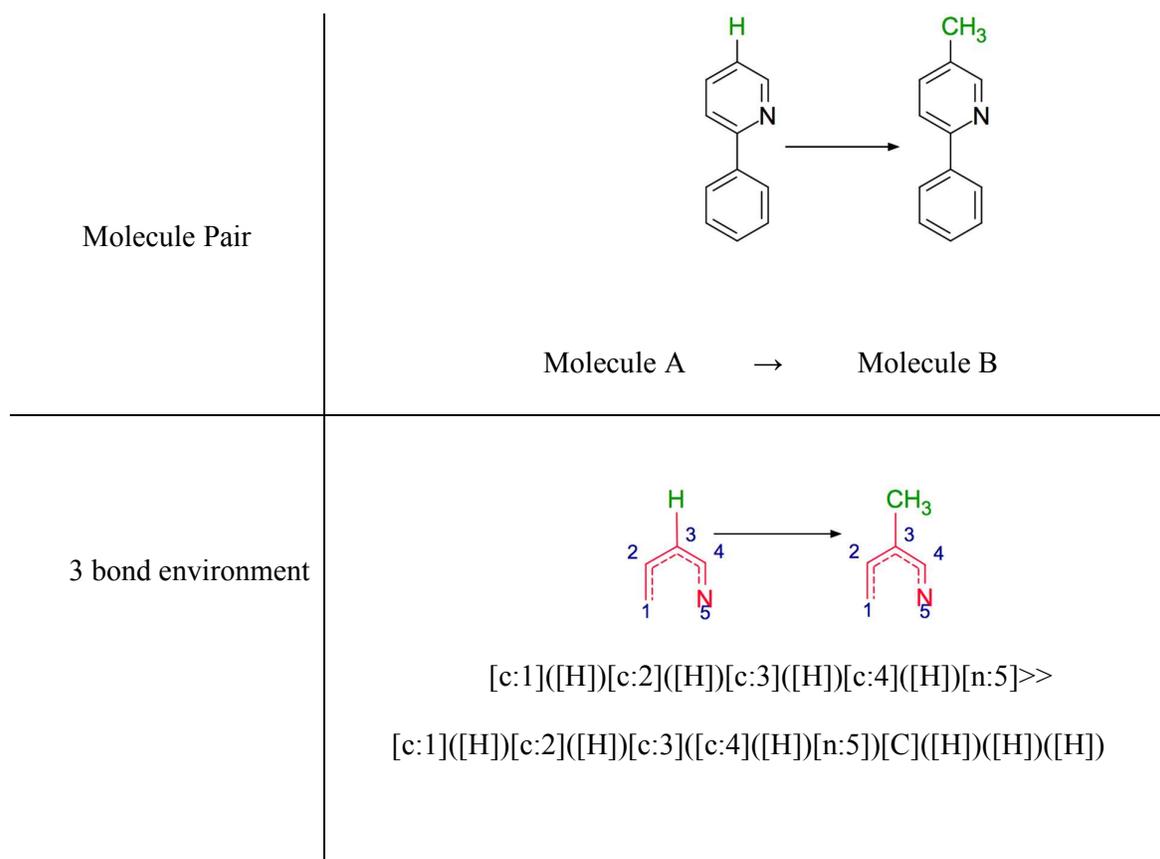
We used logD_{7.4}, solubility, in vitro microsomal and hepatocyte clearance (human, rat, mouse, cynomolgus monkey, dog), Madin Darby Canine Kidney cells (MDCK cells) permeability (A-B, B-A and efflux), cytochrome P450 inhibition (2C9, 2D6, 3A4, 2C19, 1A2), NaV 1.5 and human ether-a-go-go related gene product (hERG) ion channel inhibition, glutathione stability, and PPB (human, rat, mouse, cynomolgus monkey, dog) assays. We have not addressed explicitly the propagation of experimental errors, or the heteroscedasticity of assay data.¹⁶ In keeping with a “Big Data” paradigm we have chosen to include as much measured data as possible and to use non-parametric statistics to generate inferences that may be more resistant to the effects of noise.

Matched Molecular Pair (MMP) analysis

Compound structures were extracted as SMILES, desalted and had their charge corrected to the neutral form by applying SMIRKS using the Openeye OEChem toolkit. A set of SMIRKS were applied to recursively standardize tautomers to a common form. These were not intended to be completely accurate tautomeric forms for a given physical state, but to provide a unique canonical standard. For chiral compounds, racemates were excluded from the analysis. No further filters were applied to the inclusion of compounds.

An MMP analysis platform, MCPairs, was used that combines the benefits of the Hussain and Rea algorithm¹⁷ with those of the WizePairZ algorithm.¹⁸ Compounds were fragmented on all noncyclic single bonds except amide (C-N) and sulfonamide (S-N) bonds. Fragmentations of up to three cuts were created. Pairs were matched based on a joint constant

part, and the non-identical parts were taken as the transformations. The transformations were recorded as SMIRKS with explicit hydrogens to ensure correct aromaticity and tautomerization assignment. For a matched pair change, the local chemical structures up to four bonds from the point of change were captured, which we classify as the “environment” in the same manner as in WisePairZ. Note that within this formalism, two compounds can be linked by several transformations. Pairs were only considered if the part encoded in the SMIRKS did not make up more than 45% of the molecule and the number of non-hydrogen atoms in the changing part of the transformation was 16 or fewer. Pairs were not used if the measured activities for both compounds have a qualifier of the same direction, for example both compounds were measured as > X. An example for a pair is given in Figure 1. More examples can be found in the supporting information Figure S1.



1
2
3 Figure 1: Functional group transformation with environment specification of 3 bonds as
4 chemical sketch and as SMIRKS notation (examples for other environments can be found in
5 the supporting information). The green groups are those being changed in the transformation,
6 the red portions of the transformations show the increasing level of environment specification
7 increasing in all directions from the point of change. The blue atom numbers correspond to
8 the atom mappings in the SMIRKS transformation encoding.
9
10
11
12
13
14
15
16
17

18 *Software*

19
20 Data extraction, manipulation and analysis were performed using scripts written in
21 Python¹⁹ using the Openeye cheminformatics toolkit²⁰ for chemical structure manipulation.
22 MySQL²¹ was used for database infrastructure and statistical routines were implemented in
23 R²² using the rpy2²³ interface to Python.
24
25
26
27
28
29

30 *Rule merging/ calibration*

31
32
33 MMP identification was carried out within each contributing company behind their
34 firewalls. Data tables of the transformations with the particular biological or physical
35 chemistry delta values and associated calculated properties were then supplied to
36 MedChemica as a neutral third party. These data tables excluded the original compound
37 structures. The overall concept is shown in Figure 2.
38
39
40
41
42
43
44

45 Complete isolation of different companies' data was ensured by MedChemica acting as a
46 neutral third party and conducting all knowledge extraction through aggregation and filtering.
47 This ensured that there could be no "drill down" from any rule to any companies'
48 substructures or data by multiple layers of anonymisation.
49
50
51
52
53
54
55
56
57
58
59
60

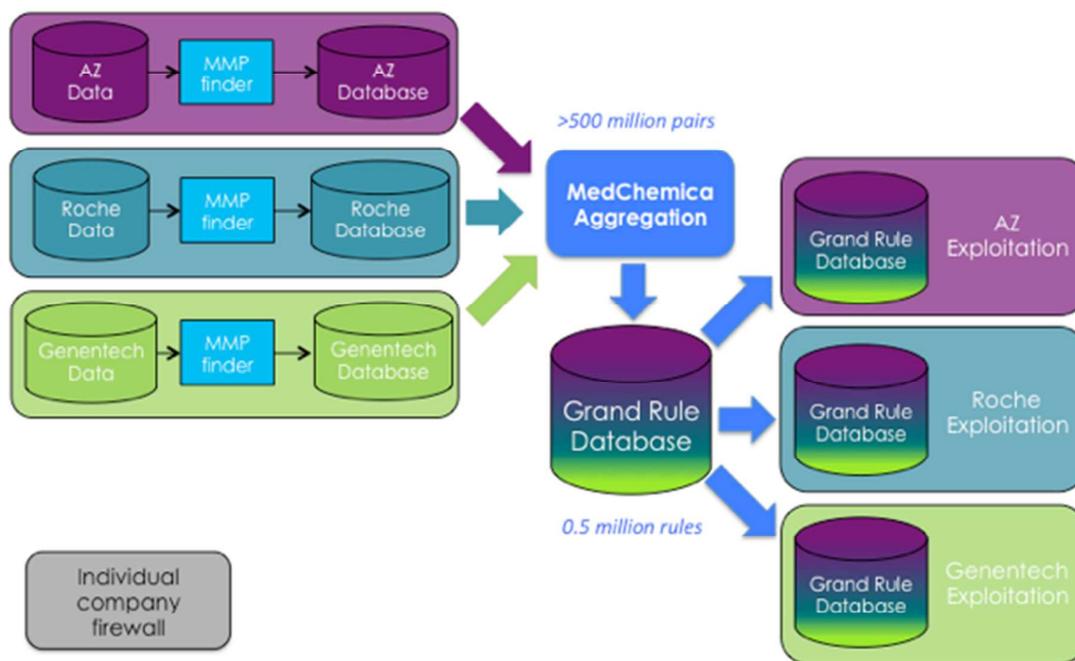


Figure 2: General assembly and use concept for the Grand Rule Database

To cover different possibilities, rule statistics were calculated in two different ways. Assuming that the differences between all pairs are normally distributed, median, standard deviation, and standard error of the mean was calculated for each transformation. Since many transformation involve censored data (*i.e.* measured values with “>” or “<” signs), non-parametric statistics based on a two-tailed binomial test were calculated to detect whether a given transformation significantly increases or decreases a property. Since the two-tailed binomial test requires at least six pairs for significance at the $p=0.05$ level, all rules with less than six pairs were discarded. An additional feature of this approach is that chemical transformations that have only been tried very few times will not be shared amongst contributing parties. There is a tension in this choice in that these may be “emerging” novel medicinal chemistry approaches which would be valuable to further explore, but in contrast

1
2
3 these may be the “novel” medicinal chemistry approaches being most recently developed
4
5 within a company and for which they wish to maintain an advantage. More details on the rule
6
7 processing can be found in the supporting information.
8
9

10
11 When only one company contributed to an ADMET endpoint, the rules extracted by this
12
13 process were added to the grand rule database (GRD). Where biological or physical
14
15 chemistry endpoints could be compared between companies, the statistically significant
16
17 transformations (according to the two-tailed binomial test) were extracted for calibration
18
19 from each company set, and then the intersection set of transformations (those with identical
20
21 canonical SMIRKS) was found. This intersection set was used to define calibration factors
22
23 between the assays using a two-way Analysis of Variance (ANOVA) procedure.²⁴ A
24
25 minimum of 8 transformations in the intersection set was required for calibration. Each of the
26
27 transformations required a minimum of 6 matched pairs. If there were more than 200
28
29 transformations in the intersection set, a random subset of 200 transformations was used. As
30
31 all transformations are mirrored, ($A \rightarrow B$ and $B \rightarrow A$ directions) for any given pair of
32
33 transformations, only the positive median direction transformation was used to avoid a
34
35 significant overestimation of the correlation between contributors as obviously $\Delta A \rightarrow B =$
36
37 $-1 * \Delta B \rightarrow A$. The mean change for the set of transformations, the means by contributor
38
39 and the grand mean coefficients were taken from the ANOVA coefficients. The calibration
40
41 factors for each contributor were calculated as the ratio of the grand mean over contributor
42
43 mean. Each contributors data set was then scaled by the calibration factor and re-aggregated
44
45 to generate a ‘common calibrated value’ representing the best estimates for each
46
47 transformation. Merging all pairs with adjusted differences into rules gave a significant
48
49 synergy benefit, where for example a transformation that had only 3 examples from one
50
51 company, 2 from another and 1 from a third company could now pass the binomial test if all
52
53
54
55
56
57
58
59
60

the examples changed the property in the same direction. The scaling procedure is schematically shown in figure 3.

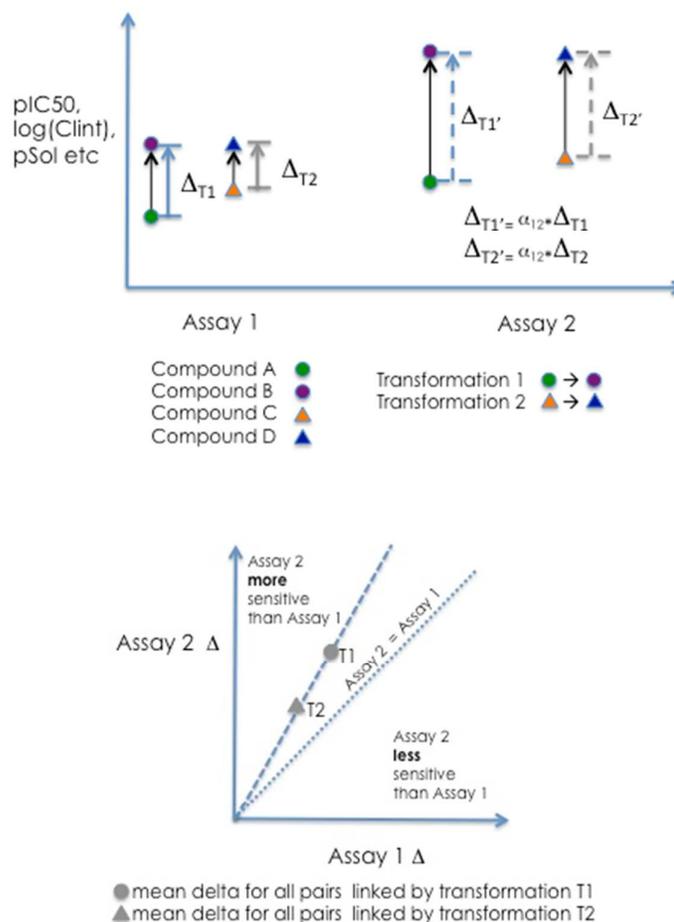


Figure 3: Linearly displaced assays of different sensitivity

The statistical methods appropriate for this comparison have been developed in the clinical chemistry arena to allow different laboratories to compare their results against each other²⁵ and more generally in the meta-analysis of clinical trials.²⁶ This is equivalent to the established medicinal chemistry practice of recording the ranking of changes in extracting knowledge from publications or presentations; the exact values might be different but the ranking should be the same. It is reasonable to assume that scientists constructing assays in

different organizations will have developed them such that they give the same ranking of standard compounds if not an identical numeric result.

Results

Rule gain

Over all the sets of assay data analyzed, 43 million unique structural transformations were found. 372,419 of these had more than six examples in any assay and passed the binomial test, each forming a medicinal chemistry rule. The distribution and overlap in rules per company is shown in figure 4.

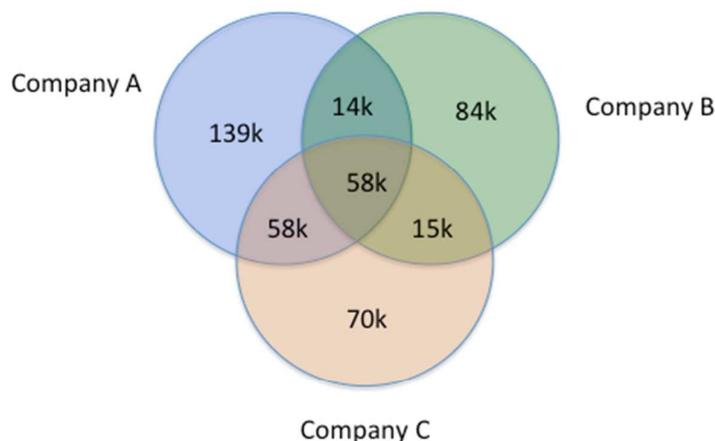


Figure 4: The origin of rules by company. The overlaps indicate the rules that contain examples from multiple companies, *i.e.* 58,000 rules had examples from all three companies, 139,000 rules were derived from company A data only.

A key question is: “what does a company gain from sharing data with the others”. In numeric terms the average gain in rules for a company is 350% - *i.e.* if a company found 100 rules from analyzing its own data sets, it would have 450 rules by joining the consortium. On

1
2
3 average this indicates how much non-overlap there has been between medicinal chemistry
4
5 practice, with a large number of rules resulting from swapping knowledge from different
6
7 areas. Also we can calculate the synergistic gain as the increase in number of rules found by
8
9 merging the pairs and analyze the combined set minus the sum of what companies would
10
11 have found individually, the average synergistic gain was 17%. This is the proportion of
12
13 rules that had too little statistical support in any individual company, but “appear” by
14
15 merging data. Some insight into the question “how large is the medicinal chemistry tool kit“
16
17 can also be addressed using this dataset, since the number of unique transformations can be
18
19 extracted. For the 372,419 transformations that have been tried enough times to be assessed,
20
21 there were 126,064 different A→B modifications when ignoring the local chemical
22
23 environment. This also suggests that medicinal chemists in different companies are working
24
25 in different areas of chemical space.²⁷⁻²⁹ Analysis of the frequency of matched pairs showed a
26
27 Zipfian distribution as previously reported by Hussain and Rea.¹⁷ An example is shown for
28
29 one company in the consortium’s contribution of human microsomal clearance data, of the
30
31 400,191 observed matched pairs, there were 303,966 unique SMIRKS. A breakdown of the
32
33 numbers of examples of SMIRKS is shown in Figure 5. 86% of the transformations were
34
35 only observed once, with only 1% being observed six or more times.
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

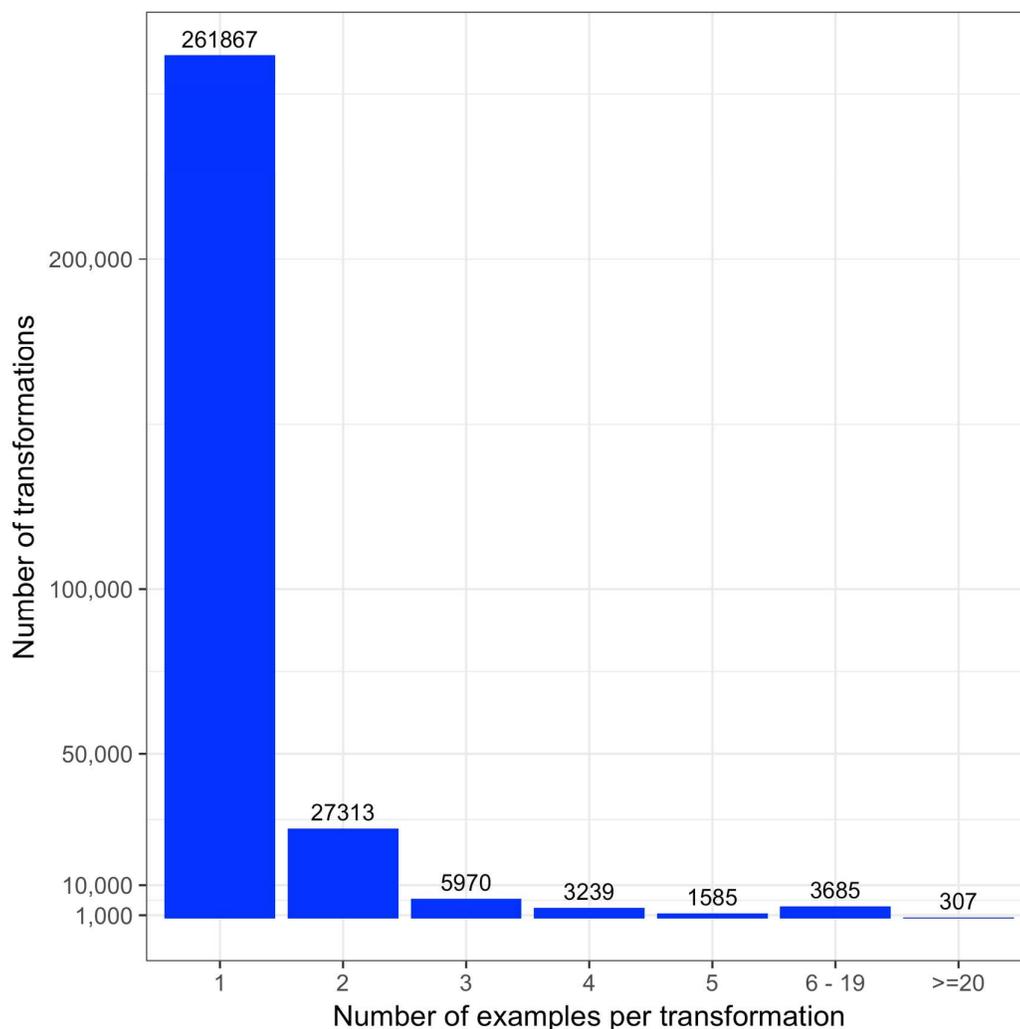
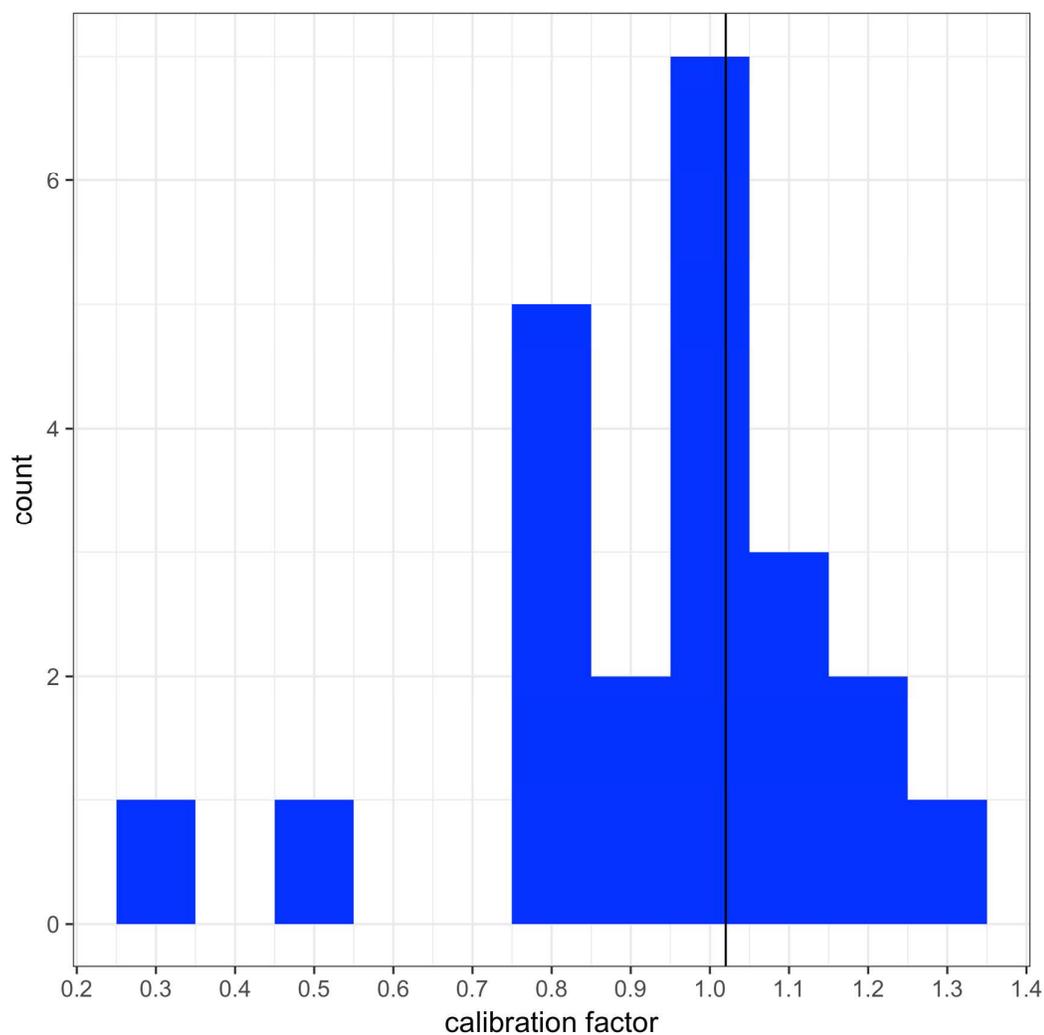


Figure 5: Distribution of number of pairs per transformation for a single company's human microsomal clearance data

Scaling factors

The distribution of calibration factors from all assay comparisons across all properties is shown in Figure 6. The calibration factors showed a median of 1.02; 80% of the calibration factors lie between 0.78 and 1.15 with two of the low outliers being assays where different technologies were used for signal detection. This suggests that, with appropriate care, the combination of data from multiple sources is feasible and that for most properties, different

1
2
3 companies operating the same methods for a property determination generate comparable
4
5 changes in measured values.
6
7

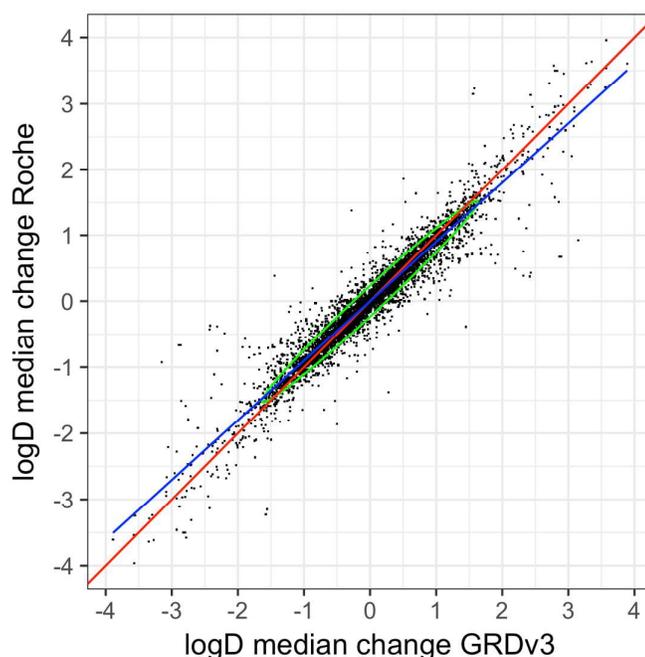


44 Figure 6: Distribution of the calibration statistics (median shown in black), count
45 indicating the number of assays for a given bin.
46
47

48 ***Roche rules vs full dataset***

49
50
51
52 One of the key questions for merging data between companies is how well these data
53 agree with one another, or in other words, if and how much noise is added by merging the
54 data. In our approach, we only formed pairs between compounds that have been measured in
55
56
57
58
59
60

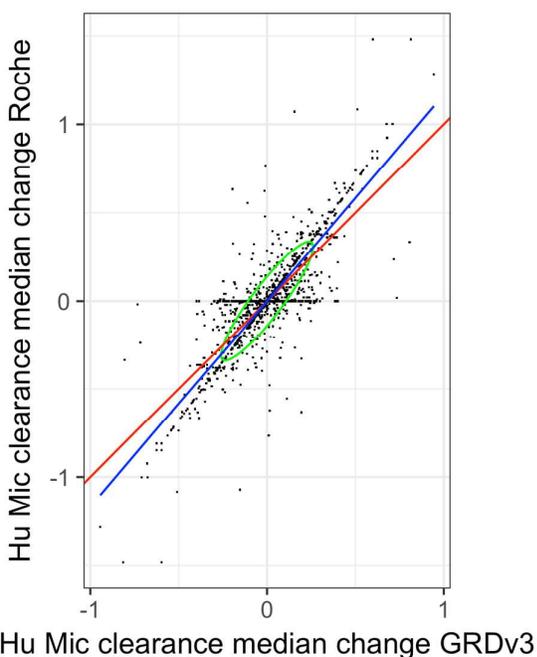
1
2
3 the same assay. It has been demonstrated that forming pairs between compounds that have
4 been measured in assays that formally measure the same endpoint but are run by different
5 operators in different places could increase the uncertainty of the rules dramatically, and the
6 number of additional pairs gained is unlikely to compensate for the noise introduced.¹⁵ For
7 GRD, we compared the rules gained from a single company database (Roche as an example)
8 versus the GRD rules (see figure 7) to compare the rules before and after the merging.
9
10
11
12
13
14
15
16



38
39 Figure 7: Correlation between predicted changes in logD based on Roche only rules and
40 joint companies' rules. Overall R^2 is 0.98. Unless stated otherwise, in this and all succeeding
41 scatterplots the pale green density ellipse contains 99%. The red line is the line of slope 1,
42 intercept 0. The blue line indicates the linear fit. In this case the vast majority of the data lies
43 on the 1:1 line between median logD change of -2 and +2 making the density ellipses hard to
44 see.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 For logD, the rules derived from Roche only data and the rules derived from the joint set
4 of pairs compare extremely well with an R^2 of 0.98 and an RMSE 0.09. Overall, there are
5 255k logD rules in the GRD and 81,617 rules based on Roche data alone.
6
7
8

9
10 LogD is a relatively simple property to measure, depending on the equilibrium of a
11 compound between two different solvents. Other important properties are more complicated,
12 and their reproducibility and transferability between companies is likely lower. In Figure 8
13 we show the correlation between rules for human microsomal clearance derived from Roche
14 data alone and the joint companies' data.
15
16
17
18
19
20
21



44
45
46
47
48
49
50

Figure 8: Comparison between Roche-only and GRD rules for human microsomal clearance. Overall R^2 is 0.76 and RMSE 0.11.

51
52
53
54
55
56
57
58
59
60

Overall, with $R^2 = 0.76$, the correlation for the microsomal clearance rules is lower than for logD. Nevertheless, almost all rules agree qualitatively, and only in 64 out of 933 the prediction changes from a negative to a positive median difference or vice versa (see Supporting Information on how positive, negative, and neutral effects are statistically defined

1
2
3 in this work). Some of the rules in GRD shown in Figure 8 are based almost exclusively on
4 Roche data. Those obviously lie close to the line of unity. For other rules, there are very few
5 pairs from Roche only, and for a subset of those rules, the Roche contributing pairs had
6 almost no change in clearance. The rules in that subset are visible as a horizontal line. Being
7 more restrictive about the inclusion criteria for the comparison, those apparent lines
8 disappear, but the overall correlation is the same (data not shown).
9
10
11
12
13
14
15
16
17

18 ***GRD rules vs previous rules***

19
20 Since the overall comparison is very favorable, we were interested in finding out how
21 individual rules from GRD compare to previously published rules. In 2015, Huchet *et. al.*
22 published a paper on fluorine effects and their impact on physicochemical properties.³⁰ From
23 prototypical examples given for logD, we extracted MMPs and compare them with the
24 statistics for the closest transformation we could find in GRD (see table 1).
25
26
27
28
29
30

31 Fluorine can have a large effect on the pKa of a nitrogen in close vicinity. Therefore,
32 transformations were selected with a balance between the number of pairs, chemical
33 similarity, and being sure that no ionizable center is affected. We did not find contextually
34 very close analogs of the compounds presented by Huchet *et. al.*. In the congeneric
35 transformation series we found, an aromatic ethoxy group is increasingly fluorinated,
36 whereas in the Huchet *et. al.* paper an aromatic propyl group is increasingly fluorinated.
37 Nevertheless, the trend reported by Huchet *et. al.* is the same as the trend we find in GRD: a
38 single fluorination reduces logD most strongly, whereas further fluorinations will reduce this
39 effect.
40
41
42
43
44
45
46
47
48
49
50

51 Source	52 Transformation	53 $\Delta\log D \pm$
		54 std (nPairs)

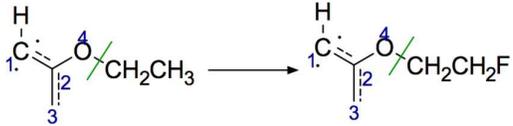
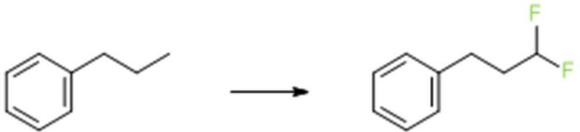
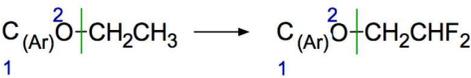
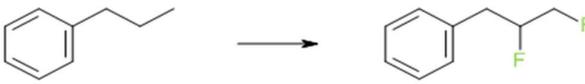
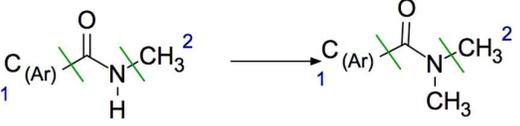
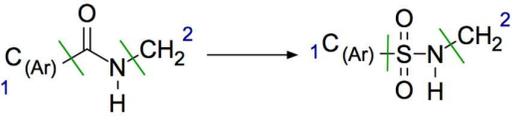
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17	Huchet		-0.7
18 19 20 21 22 23 24 25 26 27 28 29	GRD	 [c:1]([H])[c:2]([c:3])[O:4]C([H])([H])C([H])([H])([H])>>[c:1]([H])[c:2]([c:3])[O:4]C([H])([H])C([H])([H])F	-0.48 ± 0.72 (8)
30 31 32 33 34 35 36 37 38 39 40 41 42	Huchet		-0.6
43 44 45 46 47 48 49 50 51	GRD	 [c:1][O:2]C([H])([H])C([H])([H])([H])>>[c:1][O:2]C([H])([H])C([H])([F])[F]	-0.11 ± 0.28 (18)
52 53 54 55 56 57 58 59 60	Huchet		-0.4
	GRD	 [c:1][O:2]C([H])([H])C([H])([H])([H])>>[c:1][O:2]C([H])([H])C([F])([F])[F]	0.21 ± 0.59 (50)
	Huchet		-0.8 - -1.0
	GRD	insufficient examples	Insufficient examples

Table 1: Comparison between Fluorine effects published by Huchet *et. al.* and GRD statistics.

Here and in all following transformation depictions, the green bar indicates the bonds that

1
2
3 separate the constant and the variable part of the transformation. The atoms that are depicted
4 and belong to the constant part are the atoms defining the environment of the transformation.
5
6
7 As a second comparison, we looked at amide/ sulfonamide methylation. In one of the first
8
9
10 MMP analysis papers, Leach *et. al.* statistically showed that amide N-methylation increases
11
12 solubility.² Subsequently Ritchie *et. al.* presented statistics showing that on average, amide
13
14 methylation on aromatic amides reduces logD and increases solubility.³¹ They rationalize this
15
16 behavior with the planarity-breaking effect of aromatic amide N-methylation. In contrast to
17
18 amide N-methylation, they show that sulfonamide methylation always increases logD and
19
20 reduces solubility. The GRD data is in perfect qualitative agreement with their observation
21
22 (see table 2): In our data, aromatic amide N-methylation on average decreases logD by 0.25
23
24 log units and improves solubility by 0.26 log units. In contrast, aromatic sulfonamide N-
25
26 methylation increases logD by 0.37 units and decreases solubility by 0.10 log units. Note that
27
28 the rules used here are not merely hydrogen to methyl transformations, but rather specific
29
30 linker transformations as depicted in table 2.
31
32
33

Transformation	$\Delta\log D \pm \text{std}$ (nPairs)	$\Delta\text{Solubility} \pm \text{std}$ (nPairs)
 <chem>[c:1][C](=[O])[N]([H])[C:2]([H])([H])([H])>>[c:1][C](=[O])[N]([C:2]([H])([H])([H]))[C]([H])([H])([H])</chem>	-0.25 ± 0.45 (190)	0.26 ± 0.87 (144)
 <chem>[c:1][C](=[O])[N]([H])[C:2]([H])([H])([H])>>[c:1][S](=[O])(=[O])[N]([H])[C]([H])([H])([H])</chem>	-0.19 ± 0.41 (64)	-0.10 ± 0.48 (23)

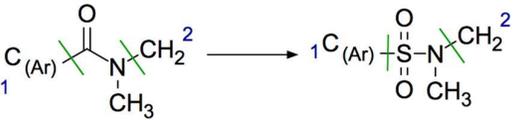
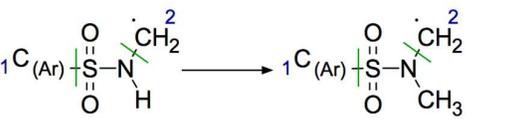
:2]([H])([H])		
 <chem>[c:1][C](=[O])[N]([C]([H])([H])([H]))[C:2]([H])([H])>>[c:1][S](=[O])(=[O])[N]([C]([H])([H])([H]))[C:2]([H])([H])</chem>	0.79 ± 0.21 (6)	insufficient examples
 <chem>[c:1][S](=[O])(=[O])[N]([H])[C:2]([H])([H])>>[c:1][S](=[O])(=[O])[N]([C]([H])([H])([H]))[C:2]([H])([H])</chem>	0.37 ± 0.35 (18)	-0.10 ± 0.67 (17)

Table 2: Effect of changing a secondary into a tertiary aromatic amide/ sulfonamide on solubility.

Another comparison with rules previously identified by Papadatos *et. al.*⁹ is detailed in the supporting information Table S17. We generally find that the rule statistics found in GRD compare very well to previously published rule statistics. Differences result from different encoding of the environment, sampling (both number of pairs and chemical diversity), and inter-company variations in measuring different endpoints (see paragraph on Roche vs. joint rules).

Solubility vs logD

The most basic physicochemical parameters of relevance for drug design are solubility and logD. Initially we explored separating the thermodynamic and kinetic solubility assays. However we discovered that the assays, although known to give different absolute values, give the same compound rankings therefore rules could be inferred from combining the mixed solubility data sets. This is a quantitative equivalent to medicinal chemistry practice

of inferring what are good “solubilizing groups” irrespective of solubility assay. LogD and solubility are often highly correlated, with the only other dominant factor apart from lipophilicity that influences solubility being the stability of the crystal structure, which can be measured by the melting point.³² A plot of the correlation between the solubility and logD rules is shown in figure 9.

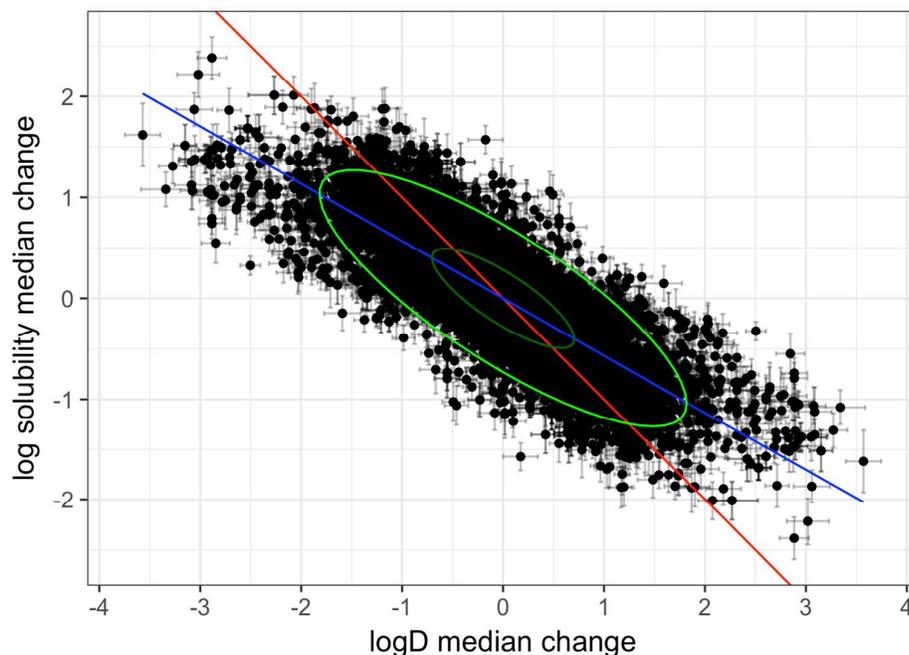


Figure 9: Solubility vs logD effects, ≥ 20 pairs per rule, $n=13453$. $R^2 = 0.66$, slope = -0.57 , intercept = 0. Red line: line of slope -1 , intercept 0. Unless stated otherwise, in this and all the succeeding scatterplots, the dark green ellipse contains 50% of the data.

Figure 9 shows that logD and solubility are also highly correlated in transformation space. The R^2 is 0.66, with a slope of -0.57 . If only rules with at least 50 pairs are considered, the R^2 increases to 0.72 (graph not shown). This means that on average a change of 1 in logD translates into a solubility change less than one order of magnitude. However the breadth of the distribution should be noted. For isolipophilic transformations, *i.e.* those with a $\Delta\log D$ of 0, the change in solubility can range over ± 1 log unit.

There are a number of rules that have good statistical support but break the logD/solubility correlation, three examples are shown in table 3.

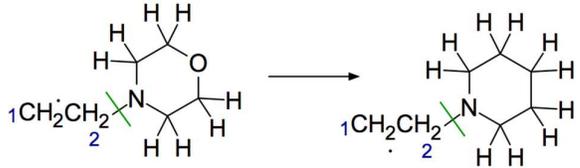
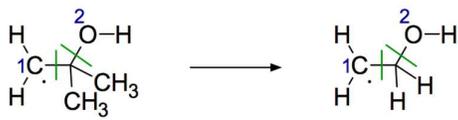
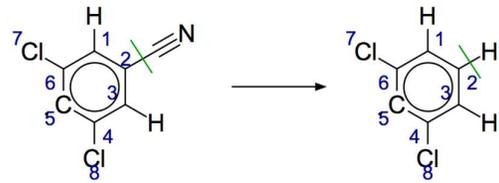
Transformation	$\Delta\log D \pm \text{std}$ (nPairs)	$\Delta\log \text{Sol} \pm \text{std}$ (nPairs)
 <chem>[C:1]([H])([H])[C:2]([H])([H])[N]1[C]([H])([H])[C]([H])([H])[O][C]([H])([H])[C]([H])([H])1>></chem> <chem>[C:1]([H])([H])[C:2]([H])([H])[N]1[C]([H])([H])[C]([H])([H])[C]([H])([H])[C]([H])([H])[C]([H])([H])1</chem>	0.00 ± 0.67 (91)	0.73 ± 0.72 (87)
 <chem>[C]([H])([H])([H])[C]([C]([H])([H])([H]))([C:1]([H])([H]))[O:2]([H])>></chem> <chem>[C:1]([H])([H])[C]([H])([H])[O:2]([H])</chem>	-0.59 ± 0.49 (82)	0.03 ± 0.72 (98)
 <chem>[c:1]1([H])[c:2]([c:3]([H])[c:4]([c:5]1[Cl:7])[Cl:8])[C]#[N]>>[c:2]1([H])[c:1]([H])[c:6]([c:5]1[Cl:8])[Cl:7]</chem>	0.45 ± 0.64 (50)	0.46 ± 1.02 (65)

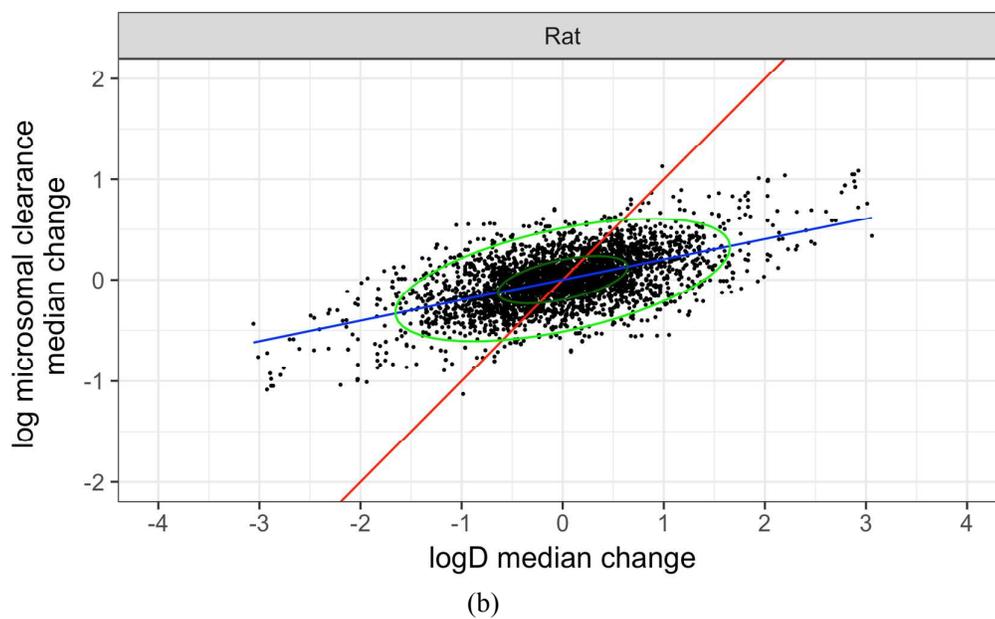
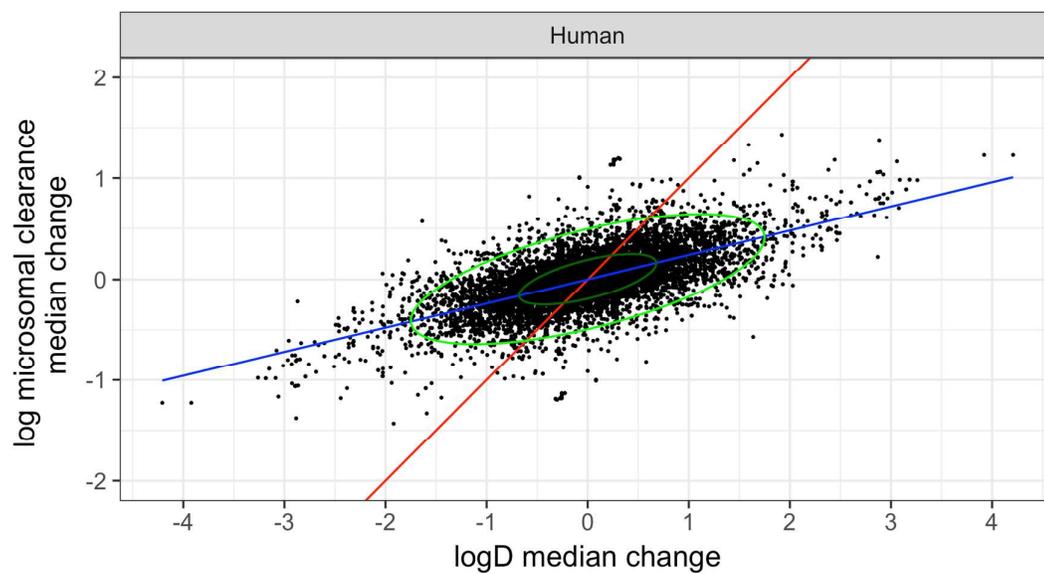
Table 3: Example transformations where the change in logD and Solubility is not as one would expect from the logD-Solubility correlation.

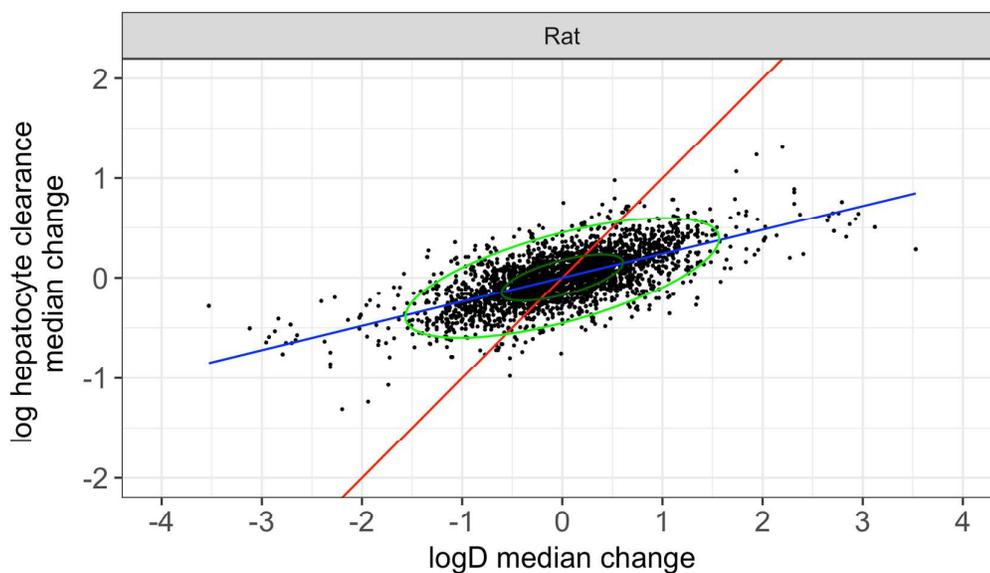
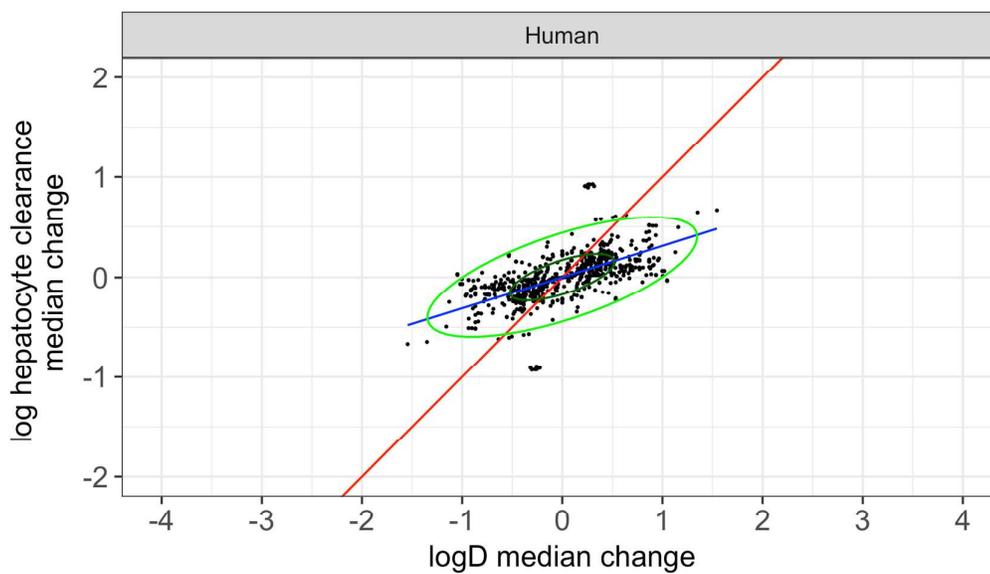
The medians for the examples shown in Table 3 are statistically all very well backed, although there is quite some variation for the individual pairs as can be seen from the standard deviations. In the first example, an aliphatically N-connected morpholine is

1
2
3 exchanged by an aliphatically N-connected piperidine.. In the second example, a tertiary
4 dimethyl alcohol is transformed into a primary alcohol. This on average reduces logD by 0.59
5 log units, but has almost no effect on solubility. In the last example, a 4-cyano-2,5-
6 dichlorophenyl is transformed into a 2,5-dichlorophenyl. This transformation increases logD
7 by 0.45 log units, but it also increases solubility by 0.46 log units. From this and other
8 examples (see supporting information) we have seen, it appears that aromatic cyano groups
9 do not help for solubility. In the supporting information, some additional transformations
10 with unexpected relationships between the effect on logD and solubility are exemplified in
11 Table S5-S7.
12
13
14
15
16
17
18
19
20
21
22
23

24 ***Clearance vs logD***

25
26 Reducing lipophilicity is often considered as a common strategy to improve metabolic
27 stability. In fact, 19 of GRD's top 20 rules (see Table S8 in supporting information) that
28 decrease liver microsomal clearance and hepatocyte clearance are also accompanied by a
29 logD decrease. Figure 10 shows a plot of the median change of microsomal and hepatocyte
30 clearance in human and rat against the corresponding median change in logD where we have
31 > 20 examples. (Microsomal and hepatocyte clearance data was used as pure Clint values,
32 uncorrected for binding.) The plot shows a good correlation between these two properties
33 ($R^2=0.40$ and 0.30 in human and rat microsomal clearance vs logD and 0.34 and 0.41 in
34 human and rat hepatocyte clearance vs logD). The plot suggests that on average a change of 1
35 in logD translates into a human or rat in-vitro clearance change of 0.2 -0.4 log units.
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60





46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 10: Median change of logD vs median change of (a) human ($n=11,572$, $R^2=0.40$, slope = 0.23) and (b) rat ($n=5,056$, $R^2=0.30$, slope = 0.20) liver microsomal clearance. Median change of logD vs median change of (c) human ($n=812$, $R^2=0.33$, slope = 0.31) and (d) rat ($n=4,937$, $R^2=0.41$, slope = 0.24) hepatocyte clearance. Data is only shown for rules based on ≥ 20 example pairs for every transformation.

There are some transformations that reduce human microsomal clearance while keeping logD neutral or even increasing it. Table 4 shows three such examples.

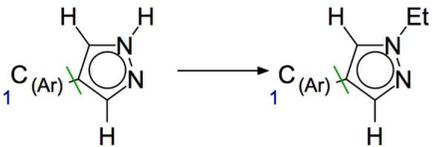
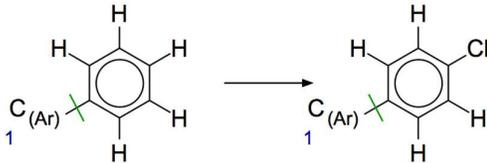
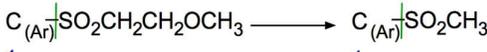
Transformation	Human microsomal Clearance median change \pm std (nPairs)	logD median change \pm std (nPairs)
 <p>[c:1][c]1[c]([H])[n]([H])[n][c]([H])1>> [c:1][c]1[c]([H])[n][n]([c]([H])1)[C]([H])([H])[C]([H])([H])([H])]</p>	<p>-0.34\pm0.71 (13)</p>	<p>0.35\pm0.45 (15)</p>
 <p>[c:1][c]1[c]([H])[c]([H])[c]([H])[c]([H])[c]([H])1>> [c:1][c]1[c]([H])[c]([H])[c]([c]([H])[c]([H])1)[Cl]</p>	<p>-0.32\pm0.51 (53)</p>	<p>0.7\pm0.74 (117)</p>
 <p>[c:1][S](=[O])(=[O])[C]([H])([H])[C]([H])([H])[O][C]([H])([H])([H])>> [c:1][S](=[O])(=[O])[C]([H])([H])([H])]</p>	<p>-0.59\pm0.38 (14)</p>	<p>0.0\pm0.11 (19)</p>

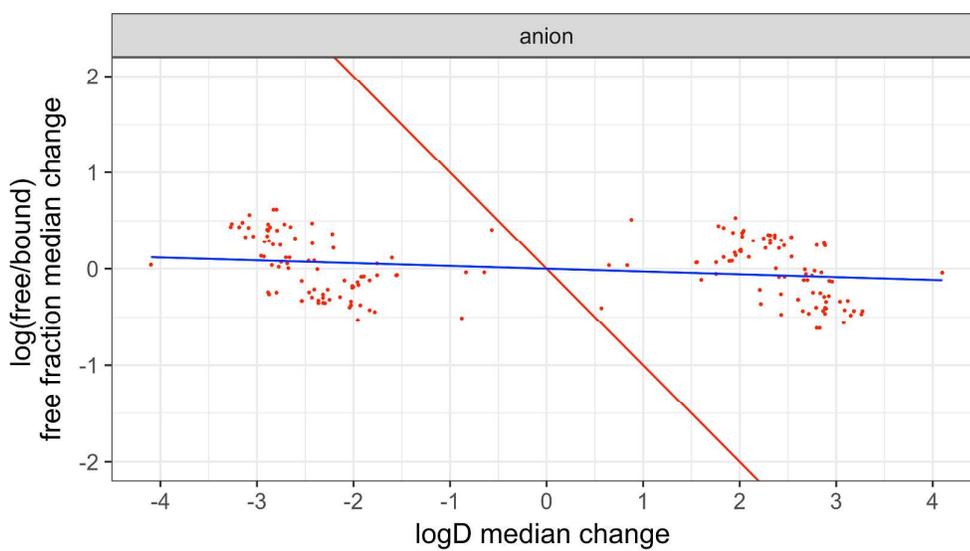
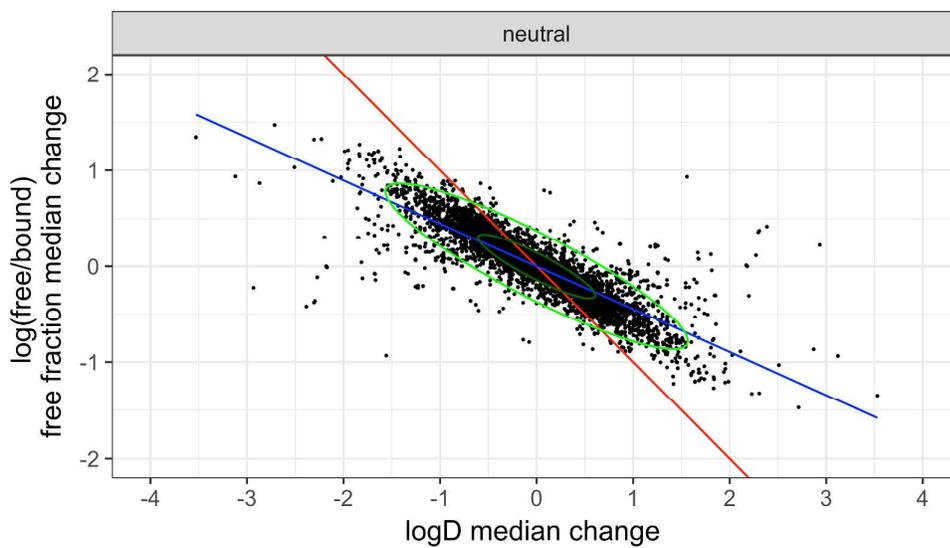
Table 4: Transformations that decrease human microsomal clearance while keeping logD constant or even increasing it.

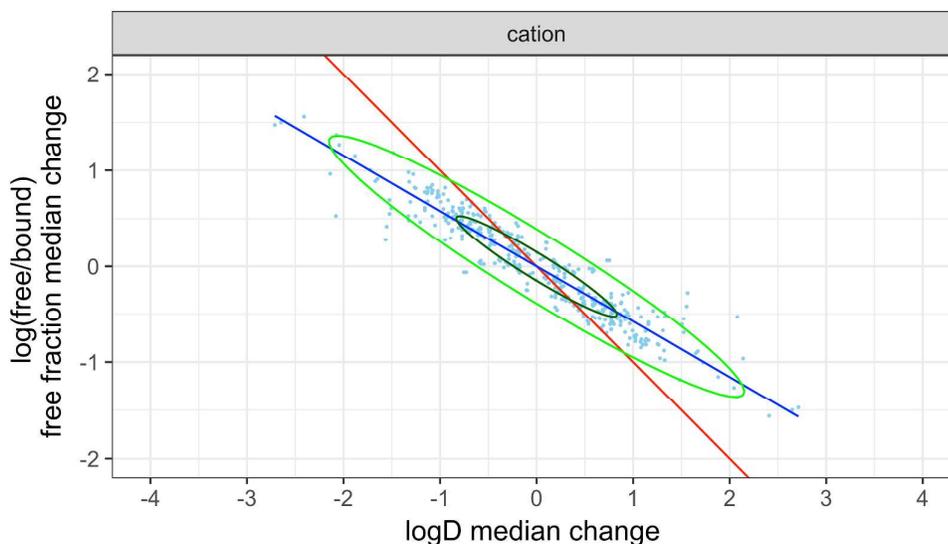
In the first transformation shown, a potentially reactive pyrazole-NH is protected by an ethyl. This increases logD by 0.34 log units, but at the same time on average reduces

1
2
3 clearance by 0.34 log units.³³ In the second example, a terminal unsubstituted phenyl ring is
4
5 substituted with chlorine in the para position. This leads to a drastic increase in logD by 0.7
6
7 log units, but also to an average decrease in clearance by 0.32 log units. Fluorine in the para
8
9 position of a phenyl ring did not produce same statistical effect as chlorine; details of chlorine
10
11 and fluorine substitution on a phenyl ring are in Table S9 in the supporting information. In
12
13 the last transformation, an aromatic methoxy-ethyl-sulfone is replaced by a methyl sulfone.
14
15 This on average does not have any effect on logD, but it strongly reduces clearance by 0.59
16
17 log units. In the supporting information (Table S10-12), more examples of transformations
18
19 that reduce clearance while keeping logD constant or even increasing it are provided.
20
21
22
23

24 ***PPB versus logD***

25
26
27 The optimization of free fraction as a goal is open to debate,³⁴ however when chosen as a
28
29 strategy, reducing logD to increase the free fraction is an accepted approach.³⁵ We expect to
30
31 observe the general trend between logD and free fraction in our data. In the GRD, changes in
32
33 logD and human free fraction correlate fairly well for rules with more than 20 examples.
34
35 Many outliers belong to the group of transformations where there is a change in anion count
36
37 (colored blue in figure 11). Those are mainly neutral to acidic or acidic to neutral
38
39 transformation changes, which have previously been reported to be “special” for free
40
41 fraction.³⁶
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



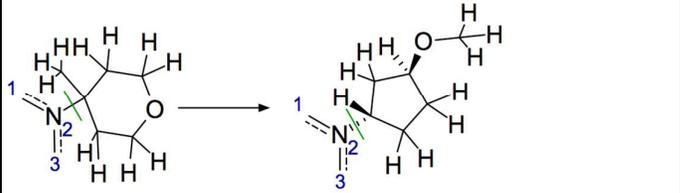
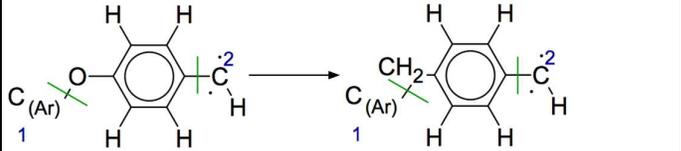
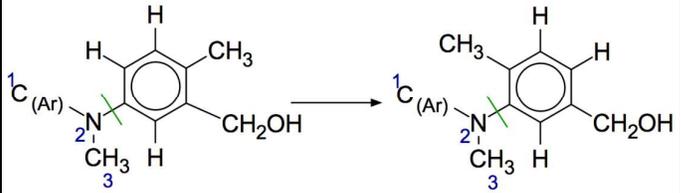


(c)

Figure 11: human free fraction vs logD effect: red line = slope -1, intercept 0. (a) Greater than or equal to 20 pairs per rule and neutral: $R^2=0.71$, Slope= -0.45, 7942 rules. (b) Rules with anion increase not equal to zero: $R^2=0.06$, Slope= -0.03, 192 rules. (c) Rules with cation increase not equal to zero: $R^2=0.90$, Slope= -0.58, 738 rules.

Figure 11 shows that differences in free fraction are highly correlated with differences in logD ($R^2=0.73$, Slope= -0.46, 8,680 rules, not visible from these plots), if no anions (*e.g.* acids) are introduced or removed. This is a valuable finding, since it means that $\Delta\log D$ can be used as a good surrogate for human $\Delta\log(\text{free}/\text{bound})$, if no exchange of anions is involved. Despite have a huge effect on logD, the addition or removal of anions on average has no effect on PPB, as can be seen from figure 11b. The change in logD due to introduction or removal of cations, in contrast, is again highly correlated to the change in PPB, as can be seen from figure 11c. If the free fraction is low, it becomes very hard to measure, and improvements in free fraction can in many cases not be monitored experimentally since they are still below the limit of detection.

Some free fraction transformations are counterintuitive because logD is either increased or constant, but the free fraction increases nevertheless, and an acid is not involved in the transformation. Table 5 shows three such examples.

Transformation	$\Delta\log D \pm \text{std}$ (nPairs)	Hu $\Delta\log$ (free/bound $\pm \text{std}$ (nPairs))
 <p>[c:1][n:2]([c:3])[C]1([C]([H])([H])[C]([H])([H])[O][C]([H])([H])[C]([H])([H])1)[C]([H])([H])([H])>> [H][C@@]1([C]([H])([H])[C]([H])([H])[C@]([C]([H])([H])1)([H])[O][C]([H])([H])([H]))[n:2]([c:1])[c:3]</p>	0.33 \pm 0.14 (11)	0.32 \pm 0.34 (11)
 <p>[c:1][O][c]1[c]([H])[c]([c]([H])[c]([c]([H])1)[C:2]([H])>> [c:1][C]([H])([H])[c]1[c]([H])[c]([H])[c]([c]([H])[c]([H])1)[C:2]([H])</p>	0.48 \pm 0.29 (11)	0.42 \pm 0.28 (12)
 <p>[c:1][N:2]([C:3]([H])([H])([H]))[c]1[c]([H])[c]([c]([H])[c]([c]([H])1)[C]([H])([H])[O]([H]))[C]([H])([H])([H])>>[c:1][N:2]([C:3]([H])([H])([H]))[c]1[c]([H])[c]([c]([H])([H])[c]([H])[c]1[C]([H])([H])([H]))[C]([H])([H])[O]([H])</p>	-0.1 \pm 0.11 (21)	0.62 \pm 0.19 (7)

1
2
3 Table 5: Transformations that increase the free fraction and increase or keep logD almost
4
5 constant.
6
7

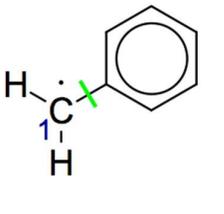
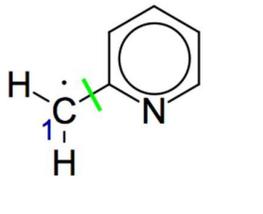
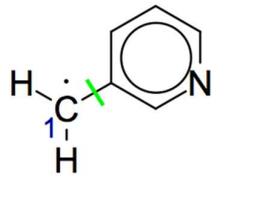
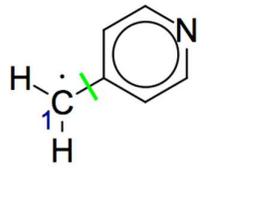
8 The first transformation in table 5 shows the conversion of a 4-methyl tetrahydropyran
9
10 into a methoxy-substituted cyclopentyl. This on average increases the free fraction by 0.32
11
12 log units, but also increases logD by 0.33 log units. The second transformation shows the
13
14 conversion of an O-linker between two aromatic rings into a CH₂-linker. LogD is increased
15
16 by 0.48 log units, while the free fraction is increased by 0.42 log units. The last
17
18 transformation in table 5 shows the shift of a methyl group on a substituted phenyl from the
19
20 para to the ortho position. LogD is not significantly changed, but the free fraction is
21
22 increased by 0.62 log units. More examples can be found in the supporting information Table
23
24 S14-S16.
25
26
27
28
29

30 ***Functional group scorecards***

31
32 Having a huge and diverse MMP database (over 350k unique structural transformations
33
34 each with more than 6 example pairs) such as the GRD allows for a characterization of
35
36 functional group replacements across many ADMET attributes. For frequently applied
37
38 transformations such as the ones shown in Tables 6 and 7, almost complete “functional group
39
40 scorecards” can be derived to predict the effect of a given frequent transformation on
41
42 ADMET and PhysChem properties, normalized to a common starting structure. The
43
44 scorecards give a quick overview of functional group properties across many different
45
46 attributes relevant to drug design. Differences in the behavior of the pyridine and chloro-
47
48 phenyl isomers is, as we might anticipate, smaller for the ADMET attributes where the
49
50 interaction is less specific such as solubility, logD and protein binding, and larger for the
51
52 more specific interactions such as P450 and hERG inhibition. Although we could speculate
53
54
55
56
57
58
59
60

on the reasons for the regioisomer differences, detailed exploration of the causes of these differences is beyond the scope of this work.

Phenyl to pyridine transformations

			
Property	$\Delta \pm \text{std (nPairs)}$	$\Delta \pm \text{std (nPairs)}$	$\Delta \pm \text{std (nPairs)}$
logD*	$-0.64 \pm 0.70 (75)$	$-0.92 \pm 0.72 (70)$	$-0.87 \pm 0.79 (54)$
Solubility	$1.00 \pm 0.75 (98)$	$0.84 \pm 0.80 (98)$	$0.79 \pm 0.81 (77)$
hERG	$-0.39 \pm 0.47 (57)$	$-0.36 \pm 0.44 (24)$	$0.12 \pm 0.54 (37)$
PPB human*	$0.31 \pm 0.42 (32)$	$0.73 \pm 0.42 (37)$	$0.79 \pm 0.38 (17)$
human Mic Clearance	$-0.11 \pm 0.46 (43)^*$	$-0.07 \pm 0.57 (90)$	Insufficient examples
human Hep Clearance	$-0.28 \pm 0.30 (8)$	$-0.14 \pm 0.13 (5)$	$-0.04 \pm 0.30 (7)$
CYP3A4 inhibition	$-0.31 \pm 0.49 (22)$	$0.65 \pm 0.54 (20)$	$0.96 \pm 0.67 (21)$
CYP2D6 inhibition	$-0.41 \pm 0.51 (8)$	$0.49 \pm 0.52 (7)$	$1.08 \pm 0.66 (8)$
CYP2C19	$-0.69 \pm 0.74 (3)$	$0.01 \pm 1.11 (8)$	$0.74 \pm 0.91 (6)$

inhibition			
------------	--	--	--

Table 6: Effects of changing a phenyl into a ortho-, meta-, and para -pyridine.* [R1][C] instead of [R1][C]([H])([H]) taken as environment due to paucity of pairs.

O- /m- /p- Chloro substitution on Phenyl

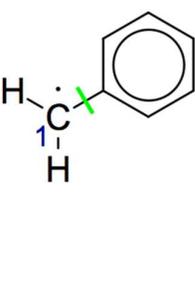
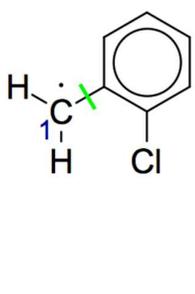
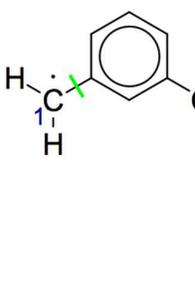
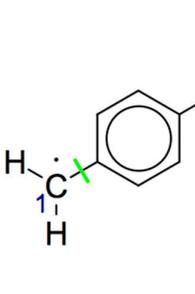
			
Property	$\Delta \pm \text{std (nPairs)}$	$\Delta \pm \text{std (nPairs)}$	$\Delta \pm \text{std (nPairs)}$
logD	$0.48 \pm 0.60 (66)$	$0.58 \pm 0.57 (86)$	$0.67 \pm 0.75 (127)$
Solubility	$-0.36 \pm 0.77 (42)$	$-0.38 \pm 0.75 (50)$	$-0.52 \pm 0.85 (81)$
hERG	$-0.04 \pm 0.29 (14)$	$0.16 \pm 0.27 (5)$	$0.38 \pm 0.35 (25)$
PPB human	$-0.33 \pm 0.34 (32)$	$-0.49 \pm 0.39 (51)$	$-0.54 \pm 0.39 (68)$
human Mic Clearance	$0.24 \pm 0.53 (47)$	$0.20 \pm 0.47 (49)$	$0.00 \pm 0.38 (61)$
human Hep Clearance	$0.16 \pm 0.38 (7)$	$0.30 \pm 0.29 (16)$	$-0.01 \pm 0.37 (20)$
CYP3A4 inhibition	$0.29 \pm 0.49 (20)$	$0.27 \pm 0.46 (22)$	$0.27 \pm 0.55 (31)$
CYP2D6 inhibition	$0.09 \pm 0.36 (9)$	$0.20 \pm 0.37 (12)$	$0.16 \pm 0.54 (24)$
CYP2C19 inhibition	$0.32 \pm 0.70 (5)$	$0.37 \pm 0.61 (8)$	$0.16 \pm 0.68 (14)$

Table 7: Effects of ortho-, meta-, and para -chloro substitution on benzyls.

Discussion

We have determined whether medicinal chemistry ADMET knowledge can be shared and enhanced by combining transformations between organizations. Using the MMPA technology, knowledge about the effects on ADMET properties of medicinal chemistry transformations can be generated based on the joint pair datasets from all participating companies. The number of rules gained increases synergistically, since there are cases where the significance threshold to define a rule is only reached by adding all pairs linked by the same transformation from different companies.

Examining the effect of the transformations on human liver microsomal stability and LogD common to all contributing organizations to those from Roche alone shows excellent agreement. This shows that the merging data by the strategy presented in this contribution supplements inhouse rules much more than it increases the noise in the rule statistics due to mixing pairs measured at different companies. Although an exact quantitative comparison with previously published rules is not possible due to different MMPA definitions used in different publications, we observe qualitative agreement between the GRD rules and published rules. At first sight the synergistic gain may appear purely quantitative (more rules), it is important to realize, however, that the more pairs support those rules, the better they become. Therefore, ‘quantity has a quality all its own’ in MMPA. Also, if the pairs come from different companies, the structural variety in the underlying pairs most probably increases, making the rules more robust. We did not quantify this here, but we believe that in future work it would be valuable to develop a metric for quantifying the structural diversity in the underlying pairs for each rule as an indicator of robustness.

The correlation between logD and solubility in transformation space, *e.g.* the space that is relevant for medicinal chemistry optimization, has an R^2 of 0.66 and agrees very well with

1
2
3 previously reported correlations for individual compounds. As such, this experiment confirms
4
5 general basic medicinal chemistry principles. However, there are also a number of exceptions
6
7 to the general logD-solubility rule. We here presented three statistically very well supported
8
9 exceptions as examples, and many more can be found in the GRD. With MMPA-based rule
10
11 databases such as the GRD, medicinal chemists can easily access and use all the “exception”
12
13 rules in prospective design.
14
15

16
17 We also showed that clearance is correlated to logD in transformation space, although the
18
19 correlation is much weaker ($R^2 = 0.3 - 0.5$). Clearance is a topic particularly well suited to
20
21 MMPA, since beyond simple logD and exposure rules, principles to address clearance are
22
23 almost exclusively based on rules about fragments and their relative stability. Here, we
24
25 present only three rules that go counter to the accepted logD correlation. The supporting
26
27 information contains some more rules, but the most natural way to access all the rules is
28
29 through MMPA-based rule databases, since there are so many rules that it is impractical to
30
31 print them all in a traditional article. With our big dataset, novel and interesting analyses
32
33 about the correlations between species and microsomal and hepatocyte clearance can be
34
35 made, all of which are however beyond the scope of this contribution.
36
37
38
39

40
41 As a third example, we present the analysis of the correlation between logD and PPB in
42
43 transformation space. We find that the two are highly correlated, if changes in the number of
44
45 anions (mostly acids) are discarded. This is an important finding: it means that in the cases
46
47 where free fraction is very hard to measure, a first estimate can be made using a matched pair
48
49 measurement of logD and the correlation between delta logD and delta log free fraction.
50
51 MMPA was critical in establishing this relationship as normal plots of free fraction versus
52
53 logD are often compromised by the noise generated by very high or low free fraction values
54
55 and the limit of PPB detection. The real relationship between changes in logD and free
56
57
58
59
60

1
2
3 fraction could only be revealed thanks to the size of the datasets and the number of rules in
4
5 GRD.
6
7

8 Relationships between rules are also important and have been explored in the related
9
10 matched molecular series studies.³⁸ With a growing body of well-defined rules across
11
12 different properties available for the common transformations, it is possible to create
13
14 functional group scorecards, such as shown above for phenyl Cl- and N-substitution (Tables 6
15
16 and 7). These scorecards enable a quick overview of functional group properties (relative to a
17
18 joint standard) and inform the choice of functional group substitution for property
19
20 optimization. Those can form the basis of analysis across the medicinal chemistry knowledge
21
22 space that may allow induction of $A \rightarrow C$ rules from the known relationships of $A \rightarrow B$ and
23
24 $B \rightarrow C$. The recent work of Kramer¹⁰ suggests that this may be possible though the
25
26 contributing transformations need to be very well exemplified to avoid the amplification of
27
28 experimental error when combining rules. Larger datasets than the one considered here may
29
30 be needed to enable these enhancements.
31
32
33
34
35

36 Finally, such a corpus of knowledge may be used for the training of medicinal chemists.
37
38 Historically medicinal chemists have learnt through their work on projects predominantly
39
40 experientially, a time consuming and potentially unreliable route to gaining expertise.³⁹ With
41
42 the possibility of encoding medicinal chemistry knowledge in a consistent, robust, data dense,
43
44 sharable manner, there is an opportunity to accelerate and enhance medicinal chemists' skills.
45
46 Access to a reference guide to what outcomes are "reasonable" for a given chemical change
47
48 can allow chemists to set proposed compounds within the context of precedents and rank
49
50 compounds in both synthetic tractability and probability of making the desired ADMET
51
52 change.
53
54
55
56
57
58
59
60

Conclusions & Outlook

We have demonstrated that significant knowledge can be extracted from large scale, unsupervised mining of in vitro ADMET data from multiple pharmaceutical companies. Within the organizations collaborating in this work and in those who have been able to exploit the data in trial form in universities and not for profit organizations, the knowledge has contributed to solving a number of drug hunting project problems leading to several publications.⁴⁰⁻⁴² We have also shown examples of transformations that run counter to established wisdom. These may be useful tools to extricate a drug hunting team from a tight optimization corner and more generally to provide opportunities to enhance our understanding of the general SAR of the system being studied. Both of these are highly desirable goals.

It is conceivable that the ensemble of rules found through MMPA on large datasets (as large as the one studied here or even larger) could form the basis for a statistics-based reference for medicinal chemistry, an “encyclopedia of medicinal chemistry tactics”.³⁷ In this case, it matters less that the underlying pairs have not all been measured at the same place. What is really important is to maximize the size of dataset to broaden the resulting corpus of rules and the statistical meaningfulness. For PPB, we obtained promising results in an analysis based on the partitioning by environment and ion class. Further partitions by factors such as shape or electronic descriptors could also be considered, though a clear mechanistic rationale justifying every partition is necessary as the analysis could fall into the traps of “data dredging” and generate false rules from statistical artifacts.

Peter Norvig, a director of research at Google has published on the “Unreasonable Effectiveness of Data”,⁴³ the concept that with a very large quantity of data, useful knowledge can be extracted in the absence of an underlying mechanistic model. For simple in

1
2
3 vitro ADMET properties it seems reasonable to suggest that we may be approaching this
4
5 point in medicinal chemistry. It is possible to view the “brute force engineering” approach of
6
7 data mining as inelegant to those more persuaded by a reductionist-theoretical approach to
8
9 compound optimization, however we see these two approaches as being on a continuum.
10
11 Established theories are often built on a modest amount of experimental data, whereas data
12
13 mining transformations brings a vast depth of data with no mechanistic model beyond the
14
15 “fundamental belief” of chemistry that “structure defines properties”. The opportunity to
16
17 enhance our SAR theories for ADMET properties by using the learning from data mining is
18
19 an area as yet under-developed and has the possibility of further accelerating drug hunting.
20
21 As researchers at Pfizer have described it, this is the process of generating tacit knowledge
22
23 from large datasets.⁴⁴ Both the reductionist-theoretical and pragmatic data mining approaches
24
25 should accelerate lead optimization. Given that the urgent and essential goal of drug
26
27 discovery is to deliver safe effective compounds to the clinic, and the continuing decrease in
28
29 drug hunting productivity⁴⁵ the simple truth is that medicinal chemists must increasingly look
30
31 to more efficient practices and reliance on robust predictions and substantiated rules.
32
33
34
35
36

37 *Supporting Information*

38
39
40
41 The supporting information contains the following information:

- 42
43
44 • Additional detail and diagrams showing the capture of the local chemical
45
46 environment round a point of change in a transformation
- 47
48 • The data merging approach
- 49
50 • Flowchart showing the method for assigning rules as significant
- 51
52 • Solubility vs logD plots and confusion tables
- 53
54 • 72 additional rules in SMIRKS and drawn transformations
- 55
56
57
58
59
60

- Comparison to 7 previously published rules
- Single company vs GRD comparison confusion tables.

Corresponding Author Information

AstraZeneca: Atilla.Ting@AstraZeneca.com +44 1625 234706

Genentech zheng.hao@gene.com +1 650 467 7446

MedChemica ed.griffen@medchemica.com +44 1625 238843

Roche christian.kramer@roche.com +41 61 6822471

Author Contributions:

The key authors contributed the work equally.

Abbreviations Used:

GRD Grand Rule Database

hERG human ether-a-go-go related gene product

MMP Matched Molecular Pair

MMPA Matched Molecular Pair Analysis

MDCK cells Madin Darby Canine Kidney cells

Acknowledgments

1
2
3 Sharing large quantities of knowledge across multiple large organizations is a rare
4 occurrence in the pharmaceutical industry and would only have been possible by the support
5 of a large technical team behind the scientists directly involved. At Genentech, Ying Littman
6 and Slaton Lipscomb provided vital technical support and Ignacio Aliagas, Alberto Gobbi,
7 Fabio Broccatelli, Jason Burch are thanked for many good discussions. We are also thankful
8 to Steve St-Gallay and John Cumming for their assistance during the early stages of the
9 collaboration. The support of the Openeye technical staff and Anthony Nicholls was highly
10 valuable in delivering the cheminformatics aspects of the collaboration. The scale of the
11 knowledge sharing undertaken would not have been possible without the support of key
12 leaders within the businesses concerned and their role in enabling such an undertaking is
13 gratefully acknowledged. In particular MedChemica is grateful to Andy Barker and Phil
14 Jewsbury for support at the genesis of this work, and Mike Snowden, Susan Galbraith, Mene
15 Pangalos and Luca Santarelli for helping prosecute the first phase of data sharing.
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33

34 *Biographies*

35 *Christian Kramer (ORCID-ID: 0000-0001-8663-5266)*

36
37
38 Christian received his PhD from the University of Erlangen, Germany, working on the
39 machine-learning prediction of physicochemical and biological properties of small molecules.
40 In 2010, he moved to Basel, Switzerland, as a Novartis Presidential PostDoc, and in 2013 he
41 accepted an assistant professorship position for Theoretical/ Computational chemistry at the
42 University of Innsbruck, Austria. In 2015, he moved to F. Hoffmann-La Roche Ltd., where
43 he works on therapeutic projects in multiple disease areas and develops computational
44 methods. Christian's research interests lie in the field of computer-aided drug design, trying
45 to map and support the complex rationalization and decision processes that underlie
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 medicinal chemistry. This includes techniques such as QSAR, QSPR, matched-molecular
4
5 pair analysis, data mining, and the analysis of experimental uncertainty.
6
7

8
9 ***Atilla Ting (ORCID-ID: 0000-0002-1590-1165)***

10
11 Atilla Ting is a Computational Chemist and a project manager working with the
12
13 Oncology Innovative Medicines drug discovery unit at AstraZeneca. Atilla's responsibilities
14
15 focus around the molecular modeling support of a wide range of early and late stage projects
16
17 along with project management responsibility. Atilla has a particular interest in structure-
18
19 based drug design and automatic matched pair analysis. Atilla has published a number of
20
21 papers in the drug discovery field. Atilla was awarded an Undergraduate degree in
22
23 Chemistry and Computer science from The University of Leeds and continued into
24
25 Postgraduate studies within the Chemistry Department to gain a PhD in Computational
26
27 Chemistry with Dr Peter Johnson before joining AstraZeneca in 2000.
28
29
30
31

32
33 ***Hao Zheng (ORCID-ID: 0000-0002-1234-1801)***

34
35 Hao Zheng is an Associate Scientist at Genentech, where he has worked for the last 11
36
37 years as a cheminformatics project manager and computational chemist in computational
38
39 drug design group. His role involves developing software solution to streamline research
40
41 workflow and provides computational modelling support to small molecule therapeutic
42
43 projects. Hao has a particular interest in applying machine learning and predictive analytics in
44
45 drug discovery. Hao holds a Masters degree in organic chemistry from Nanjing University
46
47 and a Master degree in computer science from Illinois State University.
48
49
50

51
52 ***Jérôme Hert (ORCID-ID: 0000-0003-3062-8029)***

53
54 Jérôme Hert received his PhD in cheminformatics from the University of Sheffield under
55
56 the supervision of Peter Willett. He continued to develop chemogenomics approaches as a
57
58
59
60

1
2
3 Marie Curie Fellow in the laboratories of Brian Shoichet at UCSF, and of Didier Rognan at
4
5 the University of Strasbourg. He joined F. Hoffmann-La Roche in 2009 and initially worked
6
7 on therapeutic projects in multiple disease areas. In 2013, he was appointed Head of the
8
9 Computer-Aided Drug Design section. A member of the Chemical Biology Leadership Team,
10
11 he is responsible for the coordination of cheminformatics and molecular design contributions.
12
13 Jérôme has experience and interest in the development and application of a variety of data
14
15 mining, cheminformatics, and modeling approaches. His scientific contributions include over
16
17 30 scientific articles and patents.
18
19

20 21 22 ***Torsten Schindler*** 23

24
25 Torsten Schindler studied Chemistry at the Friedrich-Alexander University in Erlangen-
26
27 Nuernberg, where he also got his PhD at the Computer Chemistry Center in the group of
28
29 Prof. Timothy Clark. From 2000 to 2002 he worked as PostDoc in the field of Computational
30
31 Chemistry at the Novartis Institute of Biomedical Research and got a permanent employee in
32
33 2002. In 2008 he moved to Basel and joined F. Hoffmann-La Roche as an Information
34
35 Analyst in Pharma Research and Early Development Informatics and is currently working in
36
37 the area of Data Science since 2015.
38
39

40 41 42 ***Martin Stahl*** 43

44
45 Martin Stahl studied chemistry in Freiburg and Würzburg, Germany, and obtained his
46
47 PhD in theoretical chemistry in Marburg with Prof. Gernot Frenking. He joined Roche in
48
49 1997. Since then, he has held various leadership roles in computational and medicinal
50
51 chemistry, biophysics and biostructure. Since 2015, he focuses on portfolio work and has led
52
53 Program Management for Small Molecule Research at Roche pRED. He is an Advisory
54
55 Board Member of Journal of Medicinal Chemistry, an editorial board member of
56
57
58
59
60

1
2
3 ChemMedChem and a Trustee of the CCDC. He is the recipient of the 2014 ACS National
4
5 Award for Computers in Chemistry and Pharmaceutical Sciences
6
7

8
9 ***Graeme Robb (ORCID-ID: 0000-0002-4531-4375)***

10
11 Graeme Robb is an Associate Principal Scientist at AstraZeneca, where he has worked for
12
13 the last 15 years as a computational chemist. He currently works within the Innovative
14
15 Medicines drug discovery unit, in the pursuit of drugs to fight cancer. His role involves using
16
17 3D molecular structure of drugs and their target proteins to simulate these systems and design
18
19 better, drug-like molecules. Graeme has been a key player in establishing numerical and
20
21 statistical learning methods for modelling and predicting the properties of molecules using
22
23 their structural features. He holds a Ph.D. and a Masters degree in chemistry from the
24
25 University of Edinburgh.
26
27
28
29

30
31 ***James J. Crawford (ORCID-ID: 0000-0002-6408-8246)***

32
33 Dr. James Crawford is a Senior Scientist and Project Team Leader in Small Molecule
34
35 Drug Discovery at Genentech. Born and raised in Glasgow, Scotland, James obtained his
36
37 MSci and Ph.D. degrees in chemistry from the University of Strathclyde, while building an
38
39 affinity for watching Rangers F.C. and listening to The Smiths. He joined Professor K.C.
40
41 Nicolaou's laboratory at The Scripps Research Institute as a Fulbright scholar, then started
42
43 his industrial career in 2006 with AstraZeneca in their Respiratory and Inflammation group.
44
45 In 2010, he moved to Genentech, where he has worked across three disease areas –
46
47 Immunology, Oncology and Antibacterials. One of his key roles at Genentech has been as the
48
49 chemistry leader of the BTK project team at Genentech.
50
51
52
53

54
55 ***Jeff Blaney (ORCID-ID: 0000-0001-9505-552X)***
56
57
58
59
60

1
2
3 Jeff received his Ph.D. in Pharmaceutical Chemistry from UCSF. He has worked in
4
5 several Pharma, Biotech, and software companies in computer-assisted drug discovery and
6
7 chemical informatics, focusing on structure-based design. He has lead Genentech's
8
9 Computational Chemistry and Cheminformatics group in Small Molecule Discovery since
10
11 October 2007. He has over 50 publications and patents.
12
13

14
15 ***Shane Montague (ORCID-ID: 0000-0003-3211-6422)***
16

17
18 Shane Montague obtained his PhD from the University of Salford, Manchester. He has
19
20 taught courses on information security and distributed systems at
21
22 undergraduate and postgraduate level. He undertook postdoctoral research in European
23
24 projects and has collaborated with international partners from across industry and academia
25
26 in both the finance and energy sectors. In 2012, he joined MedChemica Limited as a
27
28 computer scientist. He has implemented tools and systems that support the decision-making
29
30 of medicinal chemists in drug discovery. His interests lie in the area of data-intensive
31
32 applications, ranging from theory to design to implementation.
33
34
35

36
37 ***Andrew G Leach (ORCID-ID: 0000-0003-1325-8273)***
38

39
40 Andrew Leach obtained his PhD from the University of Cambridge and undertook post-
41
42 doctoral research at UCLA, supported by the Fulbright scheme. He returned to the UK to
43
44 join AstraZeneca as a computational chemist and worked in oncology, diabetes and
45
46 obesity. In 2012, he became one of the co-founders of MedChemica limited and also joined
47
48 Liverpool John Moores University as a lecturer in the School of Pharmacy and Biomolecular
49
50 Sciences. His interests include all aspects of computation to support drug discovery and
51
52 medicinal chemistry as well as the use of quantum mechanics to unravel interesting reaction
53
54 mechanisms. He has published more the 70 articles, 5 book chapters and is a named inventor
55
56 on 12 patents.
57
58
59
60

Al Dossetter (ORCID-ID: 0000-0002-4181-3193)

Al Dossetter gained his PhD from Nottingham University and after post-doctoral research at Harvard University joined AstraZeneca. He has 13 years of experience in medicinal chemistry spread across oncology (hormonal and kinase inhibitors), inflammation (OA and RA, enzyme inhibitors and GPCR targets) and diabetes (obesity, GPCR and enzyme inhibitors). In 2012, he co-founded MedChemica Limited to use Matched Molecular Pair Analysis to accelerate medicinal chemistry. MedChemica now licenses a suite of software tools for companies to extract and share knowledge from their own data and combine with public data. The software and methodologies have been used by many pharmaceutical companies, universities and biotechs to accelerate drug discovery programmes. Al is an enthusiastic advocate for his science and is frequently invited to present his research.

Ed Griffen (ORCID-ID: 0000-0003-0859-554X)

Ed Griffen obtained his PhD from Imperial College, London and undertook post-doctoral research at the University of Waterloo-Kitchener, Canada. He joined Zeneca Pharmaceuticals in medicinal chemistry working in the CNS, infection, oncology and chemical biology areas. Taking a secondment into the computational chemistry group he co-developed matched molecular pair tools to quantify medicinal chemistry approaches. In 2012 he co-founded MedChemica Ltd, a company dedicated to improving medicinal chemistry practice. He has taught medicinal chemistry courses at the University of Manchester and AstraZeneca courses in the UK, France and India. He is a named inventor on 16 patents and co-authored more than 15 articles, book chapters and a textbook. His interests are in developing data driven methods to support decision making and medicinal chemistry education.

References

- 1
2
3 (1) Dossetter, A. G.; Griffen, E. J.; Leach, A. G. Matched Molecular Pair Analysis in Drug
4
5 Discovery. *Drug Discovery Today* **2013**, *18* (15-16), 724–731.
- 6
7 (2) Leach, A. G. Matched Molecular Pairs as a Guide in the Optimization of
8
9 Pharmaceutical Properties; a Study of Aqueous Solubility, Plasma Protein Binding and
10
11 Oral Exposure. *J. Med. Chem.* **2006**, *49*, 6672–6682.
- 12
13 (3) Wassermann, A. M.; Bajorath, J. Large-Scale Exploration of Bioisosteric
14
15 Replacements on the Basis of Matched Molecular Pairs. *Future Med. Chem.* **2011**, *3*
16
17 (4), 425–436.
- 18
19 (4) Hajduk, P. J.; Sauer, D. R. Statistical Analysis of the Effects of Common Chemical
20
21 Substituents on Ligand Potency. *J. Med. Chem.* **2008**, *51* (3), 553–564.
- 22
23 (5) Wassermann, A. M.; Dimova, D.; Iyer, P.; Bajorath, J. Advances in Computational
24
25 Medicinal Chemistry: Matched Molecular Pair Analysis: Matched Molecular Pair
26
27 Analysis. *Drug Dev. Res.* **2012**, *73* (8), 518–527.
- 28
29 (6) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Methods*
30
31 *and Principles in Medicinal Chemistry*; Oprea, T. I., Ed.; Wiley-VCH Verlag GmbH &
32
33 Co. KGaA: Weinheim, FRG, 2005; pp 271–285.
- 34
35 (7) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched Molecular Pairs as a
36
37 Medicinal Chemistry Tool: Miniperspective. *J. Med. Chem.* **2011**, *54* (22), 7739–7750.
- 38
39 (8) Matlock, M.; Swamidass, S. J. Sharing Chemical Relationships Does Not Reveal
40
41 Structures. *J. Chem. Inf. Model.* **2014**, *54* (1), 37–48.
- 42
43 (9) Papadatos, G. Lead Optimization Using Matched Molecular Pairs: Inclusion of
44
45 Contextual Information for Enhanced Prediction of hERG Inhibition, Solubility, and
46
47 Lipophilicity. *J. Chem. Inf. Model.* **2010**, *50*, 1872–1886.
- 48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 (10) Kramer, C.; Fuchs, J. E.; Liedl, K. R. Strong Nonadditivity as a Key Structure–
4 Activity Relationship Feature: Distinguishing Structural Changes from Assay
5 Artifacts. *J. Chem. Inf. Model.* **2015**, *55* (3), 483–494.
6
7
8
9
10 (11) Lewis, M. L.; Cucurull-Sanchez, L. Structural Pairwise Comparisons of HLM Stability
11 of Phenyl Derivatives: Introduction of the Pfizer Metabolism Index (PMI) and
12 Metabolism-Lipophilicity Efficiency (MLE). *J. Comput. Aided Mol. Des.* **2009**, *23*,
13 97–103.
14
15
16
17
18 (12) Dossetter, A. G.; Douglas, A.; O'Donnell, C. A Matched Molecular Pair Analysis of in
19 Vitro Human Microsomal Metabolic Stability Measurements for Heterocyclic
20 Replacements of Di-Substituted Benzene Containing Compounds - Identification of
21 Those Isosteres More Likely to Have Beneficial Effects. *Med. Chem. Commun.* **2012**,
22 3, 1164–1169.
23
24
25
26
27
28
29 (13) Gleeson, P.; Bravi, G.; Modi, S.; Lowe, D. ADMET Rules of Thumb II: A Comparison
30 of the Effects of Common Substituents on a Range of ADMET Parameters. *Bioorg.*
31 *Med. Chem.* **2009**, *17* (16), 5906–5919.
32
33
34
35
36 (14) Dossetter, A. G. A Statistical Analysis of in Vitro Human Microsomal Metabolic
37 Stability of Small Phenyl Group Substituents, Leading to Improved Design Sets for
38 Parallel SAR Exploration of a Chemical Series. *Bioorg. Med. Chem.* **2010**, *18*, 4405–
39 4414.
40
41
42
43
44
45 (15) Kramer, C.; Fuchs, J. E.; Whitebread, S.; Gedeck, P.; Liedl, K. R. Matched Molecular
46 Pair Analysis: Significance and the Impact of Experimental Uncertainty. *J. Med.*
47 *Chem.* **2014**, *57* (9), 3786–3802.
48
49
50
51 (16) Wenlock, M. C.; Carlsson, L. A. How Experimental Errors Influence Drug Metabolism
52 and Pharmacokinetic QSAR/QSPR Models. *J. Chem. Inf. Model.* **2015**, *55* (1), 125–
53 134.
54
55
56
57
58
59
60

- 1
2
3 (17) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched
4
5 Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50* (3), 339–
6
7 348.
8
9
10 (18) Warner, D. J.; Griffen, E. J.; St-Gallay, S. A. WizePairZ: A Novel Algorithm to
11
12 Identify, Encode, and Exploit Matched Molecular Pairs with Unspecified Cores in
13
14 Medicinal Chemistry. *J. Chem. Inf. Model.* **2010**, *50* (8), 1350–1357.
15
16 (19) Python Software Foundation. *Python Version 2.7.1*; Available at
17
18 <http://www.python.org>.
19
20 (20) *OEChem Toolkit*, 1.7.4.3 ed.; OpenEye Scientific Software, Santa Fe, NM. OpenEye
21
22 Scientific Software, Santa Fe, NM. <http://www.eyesopen.com>.
23
24 (21) *MySQL 5.1.60, Client 14.14*; Oracle Corporation.
25
26 (22) R Core Team. *R: A Language and Environment for Statistical Computing, Version 3*;
27
28 R Foundation for Statistical Computing: Vienna, Austria, 2015.
29
30 (23) *rpy2, R in Python*.
31
32 (24) Glantz, S. A.; Slinker, B. K. *Primer of Applied Regression & Analysis of Variance*,
33
34 2nd ed.; McGraw-Hill, Medical Pub. Division: New York, 2001.
35
36 (25) Rose, C. E.; Romero-Steiner, S.; Burton, R. L.; Carlone, G. M.; Goldblatt, D.; Nahm,
37
38 M. H.; Ashton, L.; Haston, M.; Ekstrom, N.; Haikala, R.; Kayhty, H.; Henckaerts, I.;
39
40 Durant, N.; Poolman, J. T.; Fernsten, P.; Yu, X.; Hu, B. T.; Jansen, K. U.; Blake, M.;
41
42 Simonetti, E. R.; Hermans, P. W. M.; Plikaytis, B. D. Multilaboratory Comparison of
43
44 Streptococcus Pneumoniae Opsonophagocytic Killing Assays and Their Level of
45
46 Agreement for the Determination of Functional Antibody Activity in Human
47
48 Reference Sera. *Clin. Vaccine Immunol.* **2011**, *18* (1), 135–142.
49
50 (26) *Introduction to Meta-Analysis*; Borenstein, M., Ed.; John Wiley & Sons: Chichester,
51
52 U.K, 2009.
53
54
55
56
57
58
59
60

- 1
2
3 (27) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-like
4 Chemical Space Based on GDB-17 Data. *J. Comput. Aided Mol. Des.* **2013**, *27* (8),
5 675–679.
6
7
8
9
10 (28) Ertl, P. Cheminformatics Analysis of Organic Substituents: Identification of the Most
11 Common Substituents, Calculation of Substituent Properties, and Automatic
12 Identification of Drug-like Bioisosteric Groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*
13 (2), 374–380.
14
15
16
17
18 (29) Bohacek, R. S.; McMartin, C.; Guida, W. C. The Art and Practice of Structure-Based
19 Drug Design: A Molecular Modeling Perspective. *Med. Res. Rev.* **1996**, *16* (1), 3–50.
20
21
22
23 (30) Huchet, Q. A.; Kuhn, B.; Wagner, B.; Kratochwil, N. A.; Fischer, H.; Kansy, M.;
24 Zimmerli, D.; Carreira, E. M.; Müller, K. Fluorination Patterning: A Study of
25 Structural Motifs That Impact Physicochemical Properties of Relevance to Drug
26 Discovery. *J. Med. Chem.* **2015**, *58* (22), 9041–9060.
27
28
29
30
31 (31) Ritchie, T. J.; Macdonald, S. J. F.; Pickett, S. D. Insights into the Impact of N- and O-
32 Methylation on Aqueous Solubility and Lipophilicity Using Matched Molecular Pair
33 Analysis. *Med. Chem. Commun.* **2015**, *6* (10), 1787–1797.
34
35
36
37
38 (32) Ran, Y.; Yalkowsky, S. H. Prediction of Drug Solubility by the General Solubility
39 Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41* (2), 354–357.
40
41
42
43 (33) Clay, K. L.; Watkins, W. D.; Murphy, R. C. Metabolism of Pyrazole. Structure
44 Elucidation of Urinary Metabolites. *Drug Metab. Dispos. Biol. Fate Chem.* **1977**, *5* (2),
45 149–156.
46
47
48
49 (34) Liu, X.; Wright, M.; Hop, C. E. C. A. Rational Use of Plasma Protein and Tissue
50 Binding Data in Drug Design: Miniperspective. *J. Med. Chem.* **2014**, *57* (20), 8238–
51 8248.
52
53
54
55
56
57
58
59
60

- 1
2
3 (35) Meanwell, N. A. Improving Drug Candidates by Design: A Focus on Physicochemical
4 Properties As a Means of Improving Compound Disposition and Safety. *Chem. Res.*
5 *Toxicol.* **2011**, *24* (9), 1420–1456.
6
7
8
9
10 (36) Charifson, P. S.; Walters, W. P. Acidic and Basic Drugs in Medicinal Chemistry: A
11 Perspective. *J. Med. Chem.* **2014**, *57* (23), 9701–9717.
12
13
14 (37) Meanwell, N. A. Synopsis of Some Recent Tactical Application of Bioisosteres in
15 Drug Design. *J. Med. Chem.* **2011**, *54* (8), 2529–2591.
16
17
18 (38) O’Boyle, N. M.; Boström, J.; Sayle, R. A.; Gill, A. Using Matched Molecular Series as
19 a Predictive Tool To Optimize Biological Activity. *J. Med. Chem.* **2014**, *57* (6), 2704–
20 2713.
21
22
23
24
25 (39) Rafferty, M. F. No Denying It: Medicinal Chemistry Training Is in Big Trouble:
26 Miniperspective. *J. Med. Chem.* **2016** *59* (24), 10859–10864.
27
28
29
30 (40) Jordan, A. M.; Begum, H.; Fairweather, E.; Fritzl, S.; Goldberg, K.; Hopkins, G. V.;
31 Hamilton, N. M.; Lyons, A. J.; March, H. N.; Newton, R.; Small, H. F.; Vishwanath,
32 S.; Waddell, I. D.; Waszkowycz, B.; Watson, A. J.; Ogilvie, D. J. Anilinoquinazoline
33 Inhibitors of the RET Kinase domain—Elaboration of the 7-Position. *Bioorg. Med.*
34 *Chem. Lett.* **2016**, *26* (11), 2724–2729.
35
36
37
38
39
40
41 (41) Newton, R.; Bowler, K. A.; Burns, E. M.; Chapman, P. J.; Fairweather, E. E.; Fritzl, S.
42 J. R.; Goldberg, K. M.; Hamilton, N. M.; Holt, S. V.; Hopkins, G. V.; Jones, S. D.;
43 Jordan, A. M.; Lyons, A. J.; Nikki March, H.; McDonald, N. Q.; Maguire, L. A.;
44 Mould, D. P.; Purkiss, A. G.; Small, H. F.; Stowell, A. I. J.; Thomson, G. J.; Waddell,
45 I. D.; Waszkowycz, B.; Watson, A. J.; Ogilvie, D. J. The Discovery of 2-Substituted
46 Phenol Quinazolines as Potent RET Kinase Inhibitors with Improved KDR Selectivity.
47 *Eur. J. Med. Chem.* **2016**, *112*, 20–32.
48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 (42) Colley, H. E.; Muthana, M.; Danson, S. J.; Jackson, L. V.; Brett, M. L.; Harrison, J.;
4
5 Coole, S. F.; Mason, D. P.; Jennings, L. R.; Wong, M.; Tulasi, V.; Norman, D.;
6
7 Lockey, P. M.; Williams, L.; Dossetter, A. G.; Griffen, E. J.; Thompson, M. J. An
8
9 Orally Bioavailable, Indole-3-Glyoxylamide Based Series of Tubulin Polymerization
10
11 Inhibitors Showing Tumor Growth Inhibition in a Mouse Xenograft Model of Head
12
13 and Neck Cancer. *J. Med. Chem.* **2015**, *58* (23), 9309–9333.
14
15
16 (43) Halevy, A.; Norvig, P.; Pereira, F. The Unreasonable Effectiveness of Data. *IEEE*
17
18 *Intell. Syst.* **2009**, *24* (2), 8–12.
19
20
21 (44) Keefer, C. E.; Chang, G.; Kauffman, G. W. Extraction of Tacit Knowledge from Large
22
23 ADME Data Sets via Pairwise Analysis. *Bioorg. Med. Chem.* **2011**, *19* (12), 3739–
24
25 3749.
26
27 (45) Alex, A. A.; Harris, C. J.; Smith, D. A. *Attrition in the Pharmaceutical Industry:*
28
29 *Reasons, Implications, and Pathways Forward*; Wiley: Hoboken, New Jersey, 2016.
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table of Contents Graphic

