Supplemental Information to "A new Method to compute R_{complete} enables Maximum Likelihood Refinement for Small Data Sets"

J. Luebben & T.Gruene

Proceedings of the National Academy of Sciences of the United States of America

Contents

List of Tables

List of Figures

1 Numerical Results for the various Experiments of the Main Article

The Supplemental Information describes the implementation details for the main manuscript on "A new Method to compute R_{complete} enables Maximum Likelihood Refinement for Small Data Sets". It describes which programs were used and, to as great an extent as reasonable, their parameter settings. The Supplemental Material contains the numerical values and some graphical representation supporting the results presented in the main manuscript and several of the scripts that were used to run the experimental refinement and to calculate statistical values. Throughout this work, $R1$ is always calculated from the working set of reflections, *i.e.* in cases where a test set T is present, H must be replaced with $H \setminus T$ in Eq. 1 of the main manuscript.

1.1 Data Preparation and Calculation of R_{complete}

The program `crossflaghkl`, written by TG, partitions a data set of merged data. It reads in the data file in `SHELXL HKLF3` or `HKLF4` format [1]. The size of test sets is determined by the command line $-t$. Its integer parameter sets the number of flagged reflections per file. If $|T_i|$ does not divide $|H|$, the number of unique reflections, the remaining reflections are flagged in the last file. The output are thus $\lceil |H|/|T_i| \rceil$ `hkl`–files each so that each reflection of the input data file is flagged exactly once in the total of all output files.

If the input file contains Bijvoet pairs, i.e. reflections hkl and $-h-k-l$, either both reflections are flagged or neither per output file. $|T_i|$ is counted with Bijvoet pairs counted as one reflection as they are not independent.

The structural model is refined against all data to convergence. Convergence is monitored by three values printed in the `SHELXL` log–file: The $wR2$ value, the maximum coordinate shift and the maximum shift of a U–value. The `SHELXL WIGL` command causes random parameter perturbation as described in the main manuscript. The first parameter applies to the coordinates, the second parameter applies to the isotropic or anisotropic atomic displacement parameters. If one or both parameters are negative, the random shifts will be different each time `SHELXL` is run.

The resulting parameter file, the *res*-file, serves as input *ins*–file for the calculation of R_{complete} with the following modifications

1. The parameter "-1" is added to the "`L.S.`" or "`CGLS`" command respectively, causing the reflections in the test set to be excluded from refinement. As of version 2014/8, it also causes the numerator and the denominator of Eq. 3 to be printed in the log–file. R_{complete} can be conveniently calculated by summing all the numerators and all the denominators from all log–files and computing their ratio.

2. The command `LIST 9` is added. This causes the creation of a cif–file with the observed and calculated structure factor amplitudes for all reflections contained in the test set. The resulting cif–files can be concatenated with linear scaling to the maximum amplitude F_{calc} to create a cif–file in "LIST 6" format. This can be represented as electron density map by Coot [2].

Table S1: Numerical results for the values of R_{complete}, $\langle R_{\text{free}} \rangle$ and R_{boot} depending on the set size $|T_i|$ shown in Fig. S1. Data set (5). Bootstrapping with 5000 replications. min = $\min(R_{\text{free}})$, max = $\max(R_{\text{free}})$.

$\lvert T_i \rvert$	#files	R_{complete}	$\langle R_{\text{free}} \rangle$	$\sigma(R_{\text{free}})$	$\langle R1 \rangle$	$\sigma(R1)$	min	max	R_{boot}	$\sigma(R_{\text{boot}})$
1	7800	0.1651	0.2316	0.3085	0.1424	0.0001	0.0000	9.6770	0.2316	0.0035
5	1560	0.1650	0.1788	0.0877	0.1410	0.0001	0.0127	0.9584	0.1788	0.0022
10	780	0.1650	0.1720	0.0603	0.1421	0.0001	0.0417	0.4927	0.1720	0.0021
15	520	0.1650	0.1699	0.0506	0.1421	0.0001	0.0681	0.4209	0.1699	0.0022
20	390	0.1650	0.1689	0.0426	0.1421	0.0001	0.0768	0.3483	0.1689	0.0022
25	312	0.1652	0.1676	0.0388	0.1421	0.0002	0.0807	0.3227	0.1676	0.0022
30	260	0.1652	0.1673	0.0347	0.1421	0.0002	0.0906	0.2793	0.1673	0.0022
35	223	0.1652	0.1670	0.0332	0.1421	0.0002	0.1035	0.2848	0.1670	0.0023
40	195	0.1652	0.1668	0.0292	0.1421	0.0002	0.1087	0.2611	0.1668	0.0021
45	174	0.1651	0.1664	0.0280	0.1421	0.0002	0.1012	0.2450	0.1664	0.0021
50	156	0.1654	0.1665	0.0260	0.1421	0.0002	0.1075	0.2500	0.1654	0.0015
75	104	0.1652	0.1658	0.0192	0.1423	0.0002	0.1301	0.2217	0.1658	0.0019
100	78	0.1656	0.1660	0.0165	0.1423	0.0003	0.1326	0.2051	0.1660	0.0019
150	52	0.1641	0.1637	0.0128	0.1405	0.0003	0.1396	0.1411	0.1637	0.0018
500	16	0.1661	0.1663	0.0053	0.1423	0.0007	0.1554	0.1767	0.1663	0.0013

3. As the input parameters stem from a converged refinement run, the number of refinement cycles can usually be reduced to reduce the overall computation time.

The refinement runs were computed in parallel on nine i7 quad-core computers with 2.67 GHz – 2.93 GHz. Jobs were distributed with the `parallel` script from the GNU Software Foundation. In most cases, with proper balancing between the size of the test set and the computation time, R_{complete} can be calculated on such a setup within 10–20min.

Mean values and standard deviations were calculated using either `GNU octave` or `R`. `R` was also used for the bootstrapping *e.g.* for R_{boot}.

1.2 Stability with respect to the Test Set Size

For both data sets (5) and (6'), the program parameter `-t` for the program `crossflaghkl` was set to the values listed in Tables S1 and Tables S2. The results are shown as graphs in Figs. S1 and S2.

1.3 Stability of R_{complete} with Partition

The average values presented in Tab. 4 of the main text were derived in the following way:

The structural model for data set (6') was refined with `SHELXL-2014/8` with 1000 cycles of conjugate gradient least squares. The position of side chain Lys B29 was not well defined. To avoid fluctuation due to unstable residues, it was removed. The program `crossflaghkl` was applied 20 times to data set (6') with $-t50$ to create test sets of size $|T_i| = 50$. Thus each time 131 data files were created of which in the last one only 46 reflections were flagged. R_{complete}, $\langle R1 \rangle$, and

Figure S1: Variation of R_{complete} and $\langle R_{\text{free}} \rangle$ depending on test set size $|T_i|$, data set (5). (A) R_{complete} is bascially constant over the entire range of $|T_i|$ while $\langle R_{\text{free}} \rangle$ and its standard deviation vary greatly the smaller the test set size $|T_i|$. $\sigma_{\text{est}} = R_{\text{free}}/\sqrt{2|T_i|}$ for comparison [3]. (B)$\langle R_{\text{free}} \rangle$ as in (A) with minimal and maximal values on logarithmic scale. Large outliers occur with small set size $|T_i|$.

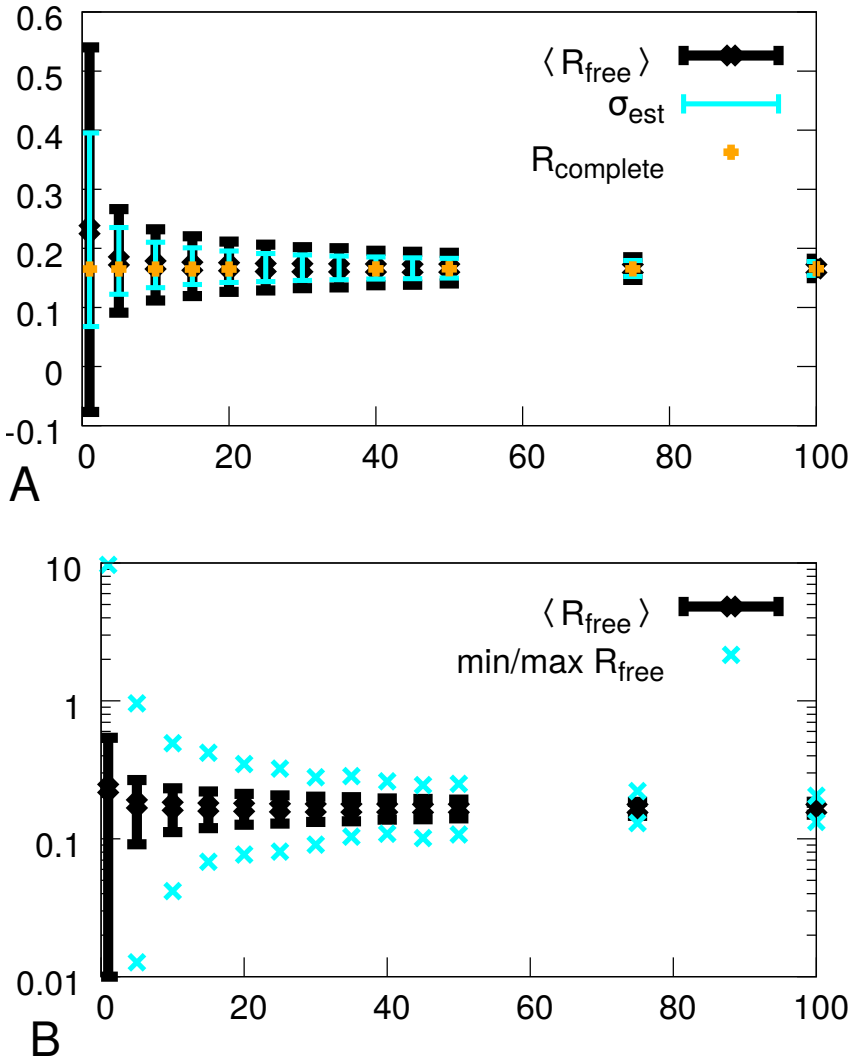


Figure S2: Same as Fig. S1 for data set (6'). The results are similar to those for data set (5).

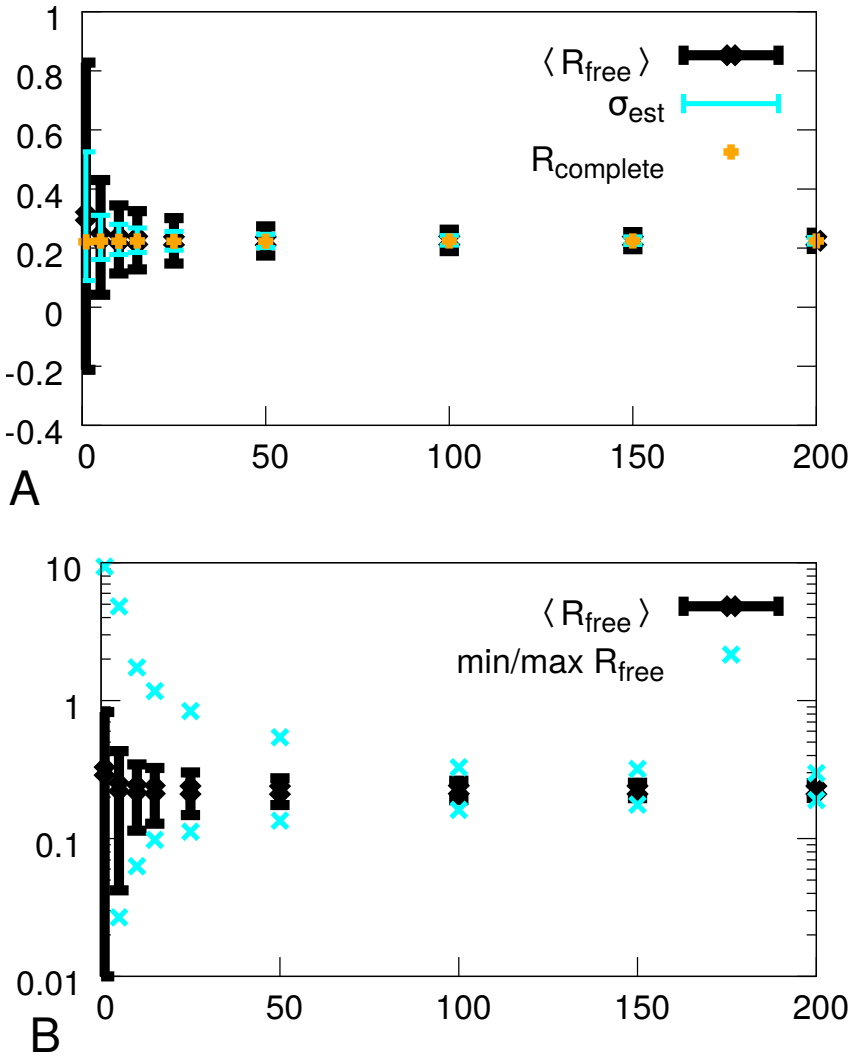


Table S2: Same as Table S1 corresponding to Fig. S2 for data set (6').

$\|T_i\|$	#files	$R_{complete}$	$\langle R_{free}\rangle$	$\sigma(R_{free})$	$\langle R1\rangle$	$\sigma(R1)$	min	max	R_{boot}	$\sigma(R_{boot})$
1	6533	0.2212	0.3079	0.5198	0.1931	0.0001	0.0000	9.3666	0.3079	0.0063
5	1307	0.2237	0.2362	0.1942	0.1941	0.0006	0.02680	4.8190	0.2362	0.0054
10	654	0.2233	0.2293	0.1153	0.1942	0.0006	0.06290	1.7404	0.2293	0.0045
15	436	0.2242	0.2265	0.0985	0.1941	0.0006	0.09780	1.1728	0.2265	0.0047
25	262	0.2230	0.2247	0.0768	0.1940	0.0006	0.11210	0.8387	0.2252	0.0047
50	131	0.2234	0.2242	0.0493	0.1941	0.0008	0.13470	0.5426	0.2242	0.0042
100	66	0.2242	0.2261	0.0364	0.1941	0.0009	0.15570	0.3316	0.2261	0.0045
150	44	0.2243	0.2251	0.0290	0.1940	0.0010	0.17610	0.3193	0.2251	0.0044
200	33	0.2241	0.2246	0.0276	0.1939	0.0011	0.19010	0.2980	0.2245	0.0047
500	14	0.2261	0.2239	0.0170	0.1934	0.0016	0.19310	0.2450	0.2239	0.0043

$\langle R_{free}\rangle$ were calculated from the values listed in the 131 `SHELXL` log–files. The average value and standard deviation were calculated with `R` from the values listed in Table S3.

1.4 Validation: How "free" is $R_{complete}$

We created two different types of experiments that allow the calculation of a bias–free R_{free} value within an acceptable amount of time and with an acceptable small amount of user interference. This allowed as to compare $R_{complete}$ with proper cross–validation and assess the bias of $R_{complete}$. We used the macromolecular data set (8) and the two small molecule data sets (2) and (3).

1.4.1 Elastase data set (8)

`crossflaghkl -t1000` created 90 data files from the merged data file for (8). Flagged reflections extracted with `GNU grep` and remove with `GNU sed`. `shelxc` [4] extracted the anomalous information from each working `hkl`-file and created the `shelxd` ins-file from the input

```
CELL 49.704    57.895    74.169    90.000    90.000    90
SPAG P212121
FIND 12
NTRY 200
SFAC S
SAD ${1}
```

`shelxe` was run from the command line with

```
shelxe kcross_set0000 kcross_set0000_fa -a -h -q
```

If the correlation coefficient between the resulting structure and the data were less than 25%, the hand needed to be inverted with the additional command line switch $-i$. The working `hkl`-file, the correct `shelxe` output PDB–file, and the amino acid sequence for porcine elastase were feed into `phenix` [5] for autobuilding. The input script was created with default options from the `phenix`

Table S3: Variation of R_{complete} with choice of partition: R-values for the structural model from data set (6') from 20 different partitionings of the data set.

run	R_{complete}	$\langle R1 \rangle$	$\sigma(R1)$	$\langle R_{\text{free}} \rangle$	$\sigma(R_{\text{free}})$
1	0.218914	0.21995	0.04882	0.19293	0.00038
2	0.219109	0.21917	0.04957	0.19293	0.00039
3	0.219369	0.22088	0.04795	0.19293	0.00036
4	0.219260	0.21983	0.04608	0.19293	0.00040
5	0.219335	0.22076	0.05326	0.19293	0.00040
6	0.219405	0.21921	0.04596	0.19293	0.00040
7	0.219059	0.22105	0.05478	0.19293	0.00039
8	0.219203	0.21955	0.04913	0.19293	0.00040
9	0.219520	0.21992	0.05056	0.19293	0.00040
10	0.219035	0.22066	0.05549	0.19293	0.00041
11	0.219133	0.22034	0.05682	0.19293	0.00044
12	0.218931	0.21983	0.04583	0.19293	0.00038
13	0.218856	0.21936	0.04695	0.19293	0.00040
14	0.219205	0.21894	0.04817	0.19294	0.00038
15	0.219282	0.21942	0.04477	0.19293	0.00040
16	0.219055	0.22055	0.05555	0.19293	0.00045
17	0.219000	0.21949	0.04287	0.19293	0.00036
18	0.219192	0.22042	0.05299	0.19293	0.00043
19	0.219632	0.22033	0.04851	0.19293	0.00041
20	0.219429	0.22143	0.05139	0.19294	0.00040

Table S4: Variation of R_{complete} with choice of partition: R-values for the structural model from data set (4) from 20 different partitionings of the data set.

run	R_{complete}	$\langle R1 \rangle$	$\sigma(R1)$	$\langle R_{\text{free}} \rangle$	$\sigma(R_{\text{free}})$
1	0.0487855	0.04917	0.00948	0.04717	0.00031
2	0.0488132	0.04924	0.00873	0.04720	0.00043
3	0.0487572	0.04935	0.01034	0.04715	0.00009
4	0.0488095	0.04932	0.00970	0.04720	0.00042
5	0.0487755	0.04928	0.00970	0.04717	0.00032
6	0.0489851	0.04944	0.00844	0.04739	0.00157
7	0.0488751	0.04933	0.01039	0.04723	0.00053
8	0.0488803	0.04937	0.00989	0.04729	0.00115
9	0.0488193	0.04890	0.00983	0.04727	0.00112
10	0.048789	0.04920	0.00883	0.04724	0.00108
11	0.0488365	0.04910	0.00948	0.04720	0.00044
12	0.048754	0.04905	0.00859	0.04717	0.00031
13	0.0488031	0.04888	0.00929	0.04720	0.00043
14	0.0488888	0.04942	0.00996	0.04724	0.00109
15	0.0487907	0.04902	0.01000	0.04723	0.00052
16	0.0490973	0.04967	0.00988	0.04742	0.00160
17	0.0490789	0.04941	0.00856	0.04745	0.00160
18	0.0488278	0.04936	0.00859	0.04727	0.00110
19	0.0489831	0.04952	0.00965	0.04751	0.00190
20	0.04878	0.04924	0.00978	0.04718	0.00032

Table S5: Variation of R_{complete} with choice of partition: R-values for the structural model from data set (7) from 20 different partitionings of the data set.

run	R_{complete}	$\langle R1 \rangle$	$\sigma(R1)$	$\langle R_{\text{free}} \rangle$	$\sigma(R_{\text{free}})$
1	0.327948	0.32853	0.03824	0.26755	0.00082
2	0.325835	0.32699	0.04560	0.26758	0.00088
3	0.325369	0.32586	0.04373	0.26753	0.00082
4	0.325369	0.32586	0.04373	0.26753	0.00082
5	0.326309	0.32745	0.04618	0.26765	0.00092
6	0.326309	0.32745	0.04618	0.26765	0.00092
7	0.325754	0.32696	0.04445	0.26771	0.00091
8	0.327227	0.32824	0.05236	0.26761	0.00106
9	0.327227	0.32824	0.05236	0.26761	0.00106
10	0.326709	0.32761	0.04420	0.26765	0.00086
11	0.326709	0.32761	0.04420	0.26765	0.00086
12	0.325287	0.32568	0.04158	0.26758	0.00081
13	0.325287	0.32568	0.04158	0.26758	0.00081
14	0.327452	0.32851	0.04889	0.26763	0.00091
15	0.326034	0.32701	0.04359	0.26764	0.00085
16	0.326034	0.32701	0.04359	0.26764	0.00085
17	0.327733	0.32806	0.04609	0.26752	0.00088
18	0.327733	0.32806	0.04609	0.26752	0.00088
19	0.326032	0.32670	0.04587	0.26758	0.00090
20	0.326032	0.32670	0.04587	0.26758	0.00090

Table S6: Mean values and standard deviations of R_{complete}, R_{free}, and $R1$ from twenty different partitions for the three data set (6'), (7), and (4). $|T_i| = 50$ in all three cases.

ID	$\langle R_{\text{complete}} \rangle$	$\sigma(R_{\text{complete}})$	$\langle R_{free} \rangle$	$\sigma(R_{\text{free}})$	$\langle R1 \rangle$	$\sigma(R1)$
(6')	0.2192	0.0002	0.2201	0.0007	0.199293	0.000003
(4)	0.0488	0.0001	0.0493	0.0002	0.0472	0.0001
(7)	0.3264	0.0009	0.3272	0.0009	0.26760	0.00005

Table S7: Comparison of R_{complete} with R_{free}, data set (8). Runs 1-30

run	$R1$	R_{free}	R_{complete}	$\langle R1 \rangle$	$\sigma(R1)$	$\langle R_{\text{free}} \rangle$	$\sigma(R_{\text{free}})$	$\langle \Delta\Phi \rangle$ [°]
01	0.3250	0.3761	0.3628	0.3249	0.0003	0.3630	0.0169	35.456
02	0.3026	0.3584	0.3381	0.3029	0.0002	0.3382	0.0160	31.280
03	0.2933	0.3234	0.3292	0.2933	0.0002	0.3293	0.0171	30.434
04	0.3102	0.3704	0.3470	0.3101	0.0002	0.3473	0.0160	32.947
05	0.3011	0.3557	0.3370	0.3010	0.0002	0.3372	0.0151	31.208
06	0.3128	0.3350	0.3514	0.3127	0.0002	0.3517	0.0165	33.455
07	0.3146	0.3692	0.3513	0.3129	0.0003	0.3515	0.0174	33.378
08	0.3142	0.3259	0.3504	0.3135	0.0002	0.3505	0.0177	33.047
09	0.3002	0.3197	0.3352	0.3000	0.0002	0.3354	0.0167	31.074
10	0.2952	0.3324	0.3307	0.2950	0.0003	0.3309	0.0161	30.575
11	0.3068	0.3512	0.3445	0.3070	0.0002	0.3448	0.0175	32.095
12	0.3248	0.3904	0.3624	0.3251	0.0002	0.3628	0.0178	34.989
13	0.2907	0.3286	0.3251	0.2903	0.0002	0.3252	0.0167	29.678
14	0.2979	0.3644	0.3344	0.2981	0.0002	0.3347	0.0159	31.076
15	0.2940	0.3197	0.3296	0.2924	0.0002	0.3298	0.0157	30.623
16	0.3223	0.3649	0.3594	0.3221	0.0002	0.3595	0.0159	35.178
17	0.3079	0.3550	0.3460	0.3081	0.0002	0.3462	0.0160	32.935
18	0.3203	0.3695	0.3576	0.3200	0.0003	0.3578	0.0170	34.497
19	0.2966	0.3361	0.3339	0.2965	0.0006	0.3342	0.0137	30.482
20	0.3044	0.3481	0.3405	0.3038	0.0003	0.3407	0.0177	31.827
21	0.2871	0.3175	0.3207	0.2864	0.0003	0.3208	0.0125	29.175
22	0.3050	0.3353	0.3419	0.3046	0.0003	0.3422	0.0160	31.868
23	0.3374	0.3723	0.3771	0.3371	0.0002	0.3773	0.0148	37.658
24	0.3071	0.3639	0.3445	0.3072	0.0002	0.3449	0.0169	31.992
25	0.2914	0.3316	0.3260	0.2914	0.0002	0.3260	0.0174	29.399
26	0.2871	0.3397	0.3235	0.2870	0.0002	0.3237	0.0156	29.362
27	0.3082	0.3331	0.3454	0.3080	0.0002	0.3455	0.0157	32.199
28	0.3166	0.3439	0.3529	0.3156	0.0002	0.3530	0.0152	33.818
29	0.2832	0.3262	0.3169	0.2817	0.0003	0.3170	0.0178	29.105
30	0.3181	0.3571	0.3569	0.3179	0.0002	0.3569	0.0145	33.554

GUI for one case. Output PDB–files did not always match and were thus superpositioned with `lsqman` [6]. PDB–files were converted to `SHELXL ins`–files with `shelxpro`. Refinement of each of the 90 structures against their respective working data set with 200 cycles of conjugate gradient least squares minimization. Bijvoet pairs were merged during refinement. The output `res` file served as input for calculating $R1$ against the respective test sets as real R_{free} free from overfitting.

The results are show in Tables S7–S9, split in 30 runs per table. $R1$ is calculated against the working set from the structural model used as input to the calculation of R_{complete}, R_{free} is calculated from the same structural model against the respective training set.

The average phase difference $\langle \Delta\Phi \rangle$ in degrees of calculated phases between the respective model and a fully refined model of Elastase was calculated with the program `phistats` [7]. The correlation coefficients between $\langle \Delta\Phi \rangle$ and $R1$, R_{free} and R_{complete} respectively were calculated with the program `GNU octave`.

Table S8: Comparison of R_{complete} with R_{free}, data set (8). Runs 31–60

run	$R1$	R_{free}	R_{complete}	$\langle R1 \rangle$	$\sigma(R1)$	$\langle R_{\text{free}} \rangle$	$\sigma(R_{\text{free}})$	$\langle \Delta\Phi \rangle$ [°]
31	0.3108	0.3462	0.3478	0.3109	0.0002	0.3478	0.0154	32.771
32	0.3208	0.3553	0.3591	0.3207	0.0002	0.3594	0.0165	34.483
33	0.3132	0.3349	0.3515	0.3131	0.0002	0.3521	0.0176	33.823
34	0.3155	0.3547	0.3523	0.3157	0.0002	0.3524	0.0164	33.162
35	0.3037	0.3626	0.3429	0.3038	0.0003	0.3431	0.0152	31.607
36	0.2863	0.3391	0.3180	0.2831	0.0003	0.3181	0.0149	28.777
37	0.3209	0.4016	0.3580	0.3213	0.0002	0.3581	0.0183	34.699
38	0.3108	0.3395	0.3459	0.3102	0.0002	0.3459	0.0169	32.490
39	0.3093	0.3604	0.3456	0.3083	0.0002	0.3459	0.0178	32.655
40	0.2932	0.3251	0.3274	0.2921	0.0003	0.3275	0.0162	30.561
41	0.2856	0.3230	0.3198	0.2854	0.0002	0.3203	0.0170	29.078
42	0.2973	0.3611	0.3337	0.2969	0.0002	0.3338	0.0156	30.444
43	0.3056	0.3337	0.3416	0.3055	0.0002	0.3418	0.0162	31.935
44	0.2980	0.3123	0.3313	0.2947	0.0003	0.3315	0.0187	30.707
45	0.3104	0.3639	0.3484	0.3104	0.0002	0.3485	0.0154	33.249
46	0.3111	0.3448	0.3475	0.3107	0.0002	0.3475	0.0165	33.154
47	0.3233	0.3870	0.3628	0.3226	0.0003	0.3629	0.0154	35.039
48	0.3336	0.4144	0.3740	0.3332	0.0003	0.3743	0.0166	36.804
49	0.2976	0.3174	0.3339	0.2966	0.0002	0.3339	0.0131	30.533
50	0.3407	0.3893	0.3809	0.3409	0.0003	0.3810	0.0177	37.634
51	0.3180	0.3536	0.3565	0.3181	0.0002	0.3567	0.0140	33.937
52	0.2997	0.3532	0.3362	0.2999	0.0002	0.3362	0.0181	31.292
53	0.3011	0.3124	0.3375	0.3009	0.0002	0.3374	0.0152	31.458
54	0.3020	0.3381	0.3376	0.3018	0.0002	0.3377	0.0161	31.266
55	0.3238	0.3626	0.3633	0.3239	0.0002	0.3634	0.0149	34.836
56	0.2941	0.3381	0.3305	0.2943	0.0003	0.3307	0.0183	30.714
57	0.3232	0.4106	0.3627	0.3233	0.0002	0.3630	0.0158	35.207
58	0.3264	0.3681	0.3659	0.3261	0.0003	0.3660	0.0163	35.515
59	0.2968	0.3426	0.3321	0.2970	0.0002	0.3322	0.0174	30.125
60	0.3182	0.3659	0.3571	0.3177	0.0004	0.3571	0.0158	34.369

Table S9: Comparison of R_{complete} with R_{free}, data set (8). Runs 61–90

run	$R1$	R_{free}	R_{complete}	$\langle R1 \rangle$	$\sigma(R1)$	$\langle R_{\text{free}} \rangle$	$\sigma(R_{\text{free}})$	$\langle \Delta\Phi \rangle$ [°]
61	0.3032	0.3338	0.3380	0.3029	0.0002	0.3384	0.0164	31.651
62	0.3180	0.3561	0.3546	0.3177	0.0004	0.3547	0.0169	33.420
63	0.3337	0.3768	0.3720	0.3338	0.0002	0.3719	0.0174	36.172
64	0.3192	0.3936	0.3590	0.3197	0.0002	0.3592	0.0158	34.692
65	0.2897	0.3503	0.3254	0.2898	0.0002	0.3255	0.0141	29.600
66	0.2881	0.3463	0.3234	0.2882	0.0002	0.3238	0.0161	29.025
67	0.3280	0.3579	0.3661	0.3281	0.0002	0.3663	0.0153	35.301
68	0.2913	0.3267	0.3275	0.2918	0.0002	0.3277	0.0176	30.040
69	0.3222	0.3750	0.3618	0.3223	0.0003	0.3621	0.0146	35.024
70	0.3002	0.3199	0.3359	0.2987	0.0002	0.3361	0.0156	31.683
71	0.2853	0.3258	0.3196	0.2848	0.0002	0.3201	0.0160	29.140
72	0.3074	0.3280	0.3460	0.3071	0.0002	0.3462	0.0157	32.168
73	0.3325	0.3739	0.3698	0.3323	0.0002	0.3702	0.0157	35.470
74	0.2957	0.3312	0.3315	0.2956	0.0002	0.3316	0.0141	30.515
75	0.2991	0.3259	0.3350	0.2990	0.0002	0.3352	0.0153	31.083
76	0.3342	0.3579	0.3717	0.3339	0.0002	0.3720	0.0154	36.362
77	0.3150	0.3609	0.3519	0.3149	0.0002	0.3520	0.0170	33.489
78	0.3066	0.3539	0.3427	0.3069	0.0002	0.3430	0.0146	32.284
79	0.3064	0.3460	0.3421	0.3063	0.0002	0.3424	0.0173	31.884
80	0.3031	0.3567	0.3431	0.3029	0.0003	0.3436	0.0173	32.433
81	0.3238	0.3651	0.3602	0.3224	0.0002	0.3604	0.0150	34.599
82	0.2929	0.3154	0.3276	0.2915	0.0002	0.3277	0.0148	30.227
83	0.3082	0.3278	0.3460	0.3078	0.0003	0.3459	0.0160	32.322
84	0.3150	0.3510	0.3523	0.3150	0.0002	0.3526	0.0167	33.583
85	0.3143	0.3623	0.3528	0.3141	0.0003	0.3529	0.0162	33.192
86	0.3139	0.3565	0.3520	0.3140	0.0003	0.3522	0.0164	33.208
87	0.3182	0.3621	0.3559	0.3182	0.0003	0.3561	0.0177	33.614
88	0.3063	0.3665	0.3455	0.3065	0.0003	0.3455	0.0152	32.374
89	0.3260	0.3766	0.3627	0.3260	0.0002	0.3628	0.0159	35.219
90	0.3147	0.3749	0.3510	0.3114	0.0005	0.3513	0.0163	33.205

Table S10: Comparison of R_{complete} with R_{free}, data set (2). The table is sorted by R_{complete} to highlight the bimodal distribution of the R_{complete} which is not apparent in the cases of $R1$ or R_{free}.

run	R_{complete}	$R1$	R_{free}
12	0.120421	0.1081	0.1322
19	0.120421	0.1094	0.1148
1	0.120422	0.1083	0.1273
13	0.120422	0.1084	0.1379
16	0.120422	0.1084	0.1352
15	0.120423	0.1088	0.1286
18	0.120423	0.1091	0.1171
10	0.120424	0.1098	0.1102
2	0.120424	0.1092	0.1101
6	0.120425	0.1096	0.1017
14	0.120426	0.1089	0.1126
7	0.120427	0.1085	0.1258
17	0.120428	0.1090	0.1166
4	0.120428	0.1088	0.1248
3	0.121181	0.1094	0.1257
8	0.121238	0.1099	0.1229
11	0.121243	0.1100	0.1223
20	0.121243	0.1096	0.1192
5	0.121245	0.1094	0.1293
9	0.121245	0.1087	0.1389

1.4.2 Small Molecule Data Sets (2) & (3)

For the data sets (2) and (3), the data were split into 20 different working/test sets. Structure solution was done with `shelxt`, refinement with `shelxl` with default settings and without any human interference. Numerical results are shown in Table S10 and Table S11.

1.5 Validation II: Comparison with Calculated Data

Data were calculated with `xprep` [1] from the two structural models for data sets (4) and (6'). The modifications mentioned in the main text were introduced by manual editing of the `SHELXL ins`-files. Map differences where computed with Coot [2]. Fig. S3 shows a stronger signal in the map computed with the R_{complete} method than the conventional map. The improvement is consistent for all the introduced modifications. Note that the resulting differences are not due to scaling errors when making the differences: otherwise density would show for the rest of the structural models; this was checked and not found, even at lower contour levels.

Results for R_{complete}, $\langle R_{\text{free}} \rangle$, and $R1$ for the calculated data from the structural models of data sets (6') and (4). $R1_{\text{ideal}}$: $R1$ value calculated after modification of the structure without any refinement. In order to check if parameter perturbation would reduce the gap between $R1_{\text{ideal}}$ and R_{complete}, we tried random parameter perturbation with an amplitude $X = 0.6$. This was large enough to move most atoms beyond the radius of convergence. One representative structural model is shown in Fig. S4 and shows why parameter perturbation must be applied with caution, if at all.

Table S11: Comparison of R_{complete} with R_{free}: Same table as Table S10 but for data set (3). In this case the sorting by R_{complete} highlights the two outlier structures in run 8 and run 1.

run	R_{complete}	$R1$	R_{free}
11	0.127292	0.1139	0.1458
4	0.127292	0.1136	0.1304
10	0.127293	0.1143	0.1204
14	0.127294	0.1135	0.1407
18	0.127294	0.1133	0.1338
3	0.127294	0.1147	0.1223
13	0.127295	0.1144	0.1242
9	0.127295	0.1134	0.1377
12	0.127296	0.1152	0.1198
16	0.127296	0.1153	0.1092
17	0.127296	0.1140	0.1326
20	0.127296	0.1151	0.1124
2	0.127296	0.1145	0.1142
5	0.127296	0.1145	0.1183
19	0.127297	0.1141	0.1307
15	0.127298	0.1146	0.1189
6	0.127301	0.1137	0.1395
7	0.127301	0.1137	0.1383
8	0.129160	0.1145	0.1460
1	0.132383	0.1196	0.1409

Figure S3: Difference between the electron density map calculated with $|F_{\text{calc}}(hkl)|$ produced by the R_{complete} method and the electron density map resulting from conventional refinement. A: Data set from (4). The red *aka* negative density indicates that the wrongly placed sodium atom has too many electrons compared to the oxygen atom contained in the ideal model. The green *aka* positive density patches nearby are most likely Fourier ripples. B: Data set from (6'). The green *aka* positive density sphere show the positions of the oxygen atoms removed from the structural model used to calculate the data. Note the cyan circles: as opposed to the left panel the difference of three electrons per sodium atoms that wrongly replaced an oxygen atom does not stand out against the eight electrons of the missing oxygens.

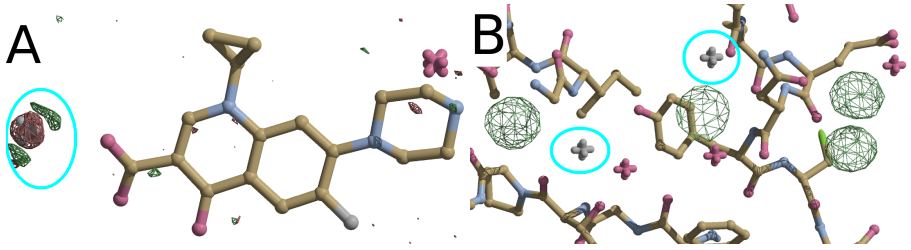


Figure S4: Random parameter perturbation with amplitude 0.6 on the structural model of data set (4) leads to destruction of the chemical integrity of the model. As discussed in the manuscript random perturbation is not necessary and not useful for bias reduction.

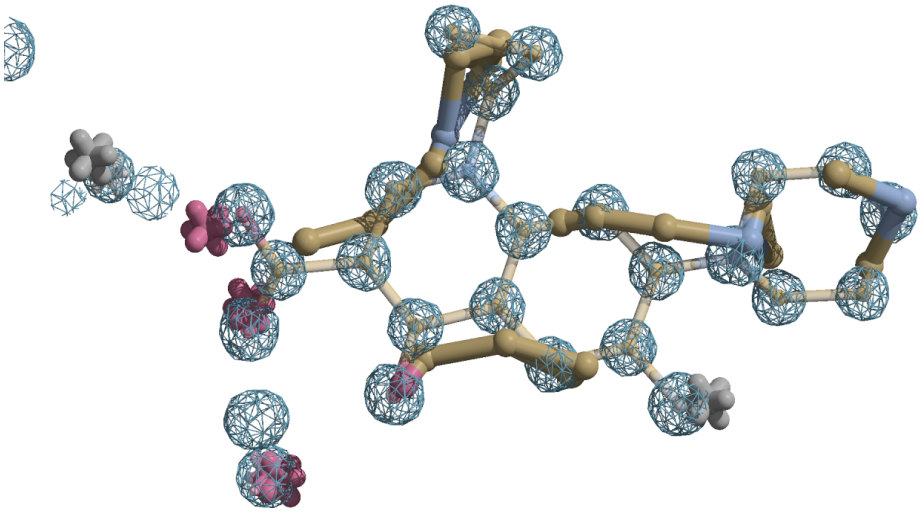
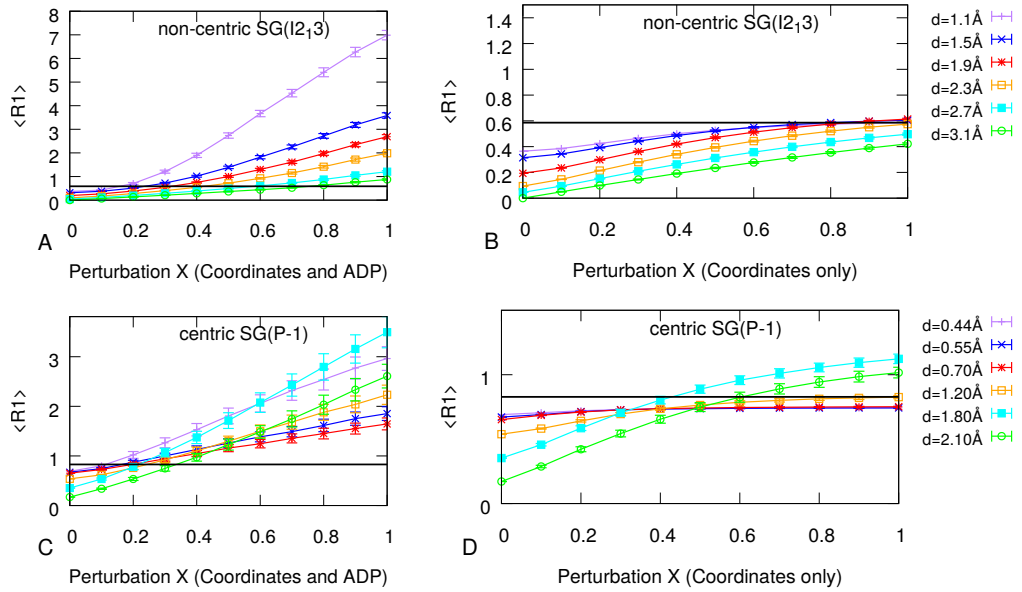


Table S12: R–values for modified structural models against calculated data. $R1_{\text{ideal}} = R1$ after modification without refinement, *i.e.* free from overfitting.

ID	$\lvert T_i \rvert$	$R1_{\text{ideal}}$	$R1$	R_{complete}	$\langle R_{\text{free}} \rangle$	$\sigma(R_{\text{free}})$
(6')	30	7.22%	6.50%	7.81%	7.97%	1.83%
(4)	10	15.30%	6.13%	6.34%	6.67%	2.65%

Figure S5: Reduction of the effect of overfitting from the structural model by perturbing all parameters (A+C) or coordinates only (B+D). $R1$ at $X = 0$ corresponding to Tables S13 and S14 respectively. A+B from the non-centric data set (6), C+D from the centric data set (1).



1.6 Effect of Parameter Perturbation

A regular grid of atoms was freely refined until convergence. Subsequently, coordinates and coordinates plus atomic displacement parameters were randomly perturbed with the commands 'WIGL -X 0' and 'WIGL -X X' respectively. The average $R1$ was calculated 500 times without further refinement. Fig. S5 and Tables S15–S26 presented the results. The parameter X was varied from 0 to 1.0 in steps of 0.1; the calculations were carried out with data cut to a resolution between 1.1 Å and 3.1 Å for data set (6) in space group $I2_13$ and between 0.44 Å to 2.1 Å for data set (1) in space group $P\bar{1}$. The phase error $\langle\Delta\phi\rangle = \arg\sum_i^N e^{i\Delta\phi_i}$ and the mean resultant length, which corresponds the the standard deviation for circular data, $\rho = \|1/N\sum_i^N e^{i\Delta\phi_i}\|$ were calculated against the phases from the unperturbed parameters. Only $\langle R1\rangle$ is displayed in the main text.

In the centrosymmetric spacegroup $P\bar{1}$, the phase angle can be either 0 ∘ or 180° and hence $\langle\Delta\Phi\rangle = 0°$ or $\langle\Delta\Phi\rangle = 180°$ or may even not be computable if $\rho \approx 0$.

1.6.1 Non–Centrosymmetric Space Group, Data Set (6)

Numerical values presented in Fig. S5 A+B are listed in Tables S15–S20.

Table S13: r and $R1$ by resolution cut–off from data set (6) to assess the reduction of the effect of overfitting from the structural model by random parameter perturbation for a non–centrosymmetric space group. $R1$: reliability index after unrestrained refinement. See Fig. S5 C+D.

$d_{\min}$	[Å]	1.10	1.50	1.90	2.30	2.70	3.10
#data		32,598	13,013	6,469	3,696	2,305	1,540
#data:#param.		6.80	2.71	1.35	0.77	0.48	0.32
R1	[%]	36.5	31.4	19.3	9.40	4.50	0.00

Table S14: Data to parameter ratio r and $R1$ by resolution cut–off from data set (1) to assess the reduction of the effect of overfitting from the structural model by random parameter perturbation for a centrosymmetric space group. $R1$: reliability index after unrestrained refinement. See Fig. S5 A+B.

$d_{\min}$	[Å]	0.44	0.55	0.70	1.20	1.80	2.10
#data		42,997	24,100	12,045	2,375	690	418
#data:#param.		80.4	45.0	22.5	4.44	1.29	0.78
R1	[%]	69.0	67.1	65.2	53.8	35.3	17.1

Table S15: Numerical Values for Fig. S5: Spacegroup $I2_13$, data set (6), $d_{\min} = 1.10$Å

	WIGL -X X			WIGL -X 0		
X	$\langle R1\rangle \pm\sigma(R1)$	$\langle\Delta\Phi\rangle$	$\langle\rho\rangle$	$\langle R1\rangle \pm\sigma(R1)$	$\langle\Delta\Phi\rangle$	$\langle\rho\rangle$
0.0	0.3645±0.0000	0.0	1.000	0.3645±0.0000	0.0	1.000
0.1	0.4287±0.0085	22.9	0.856	0.3851±0.0013	12.6	0.944
0.2	0.7038±0.0415	41.8	0.715	0.4238±0.0024	23.8	0.860
0.3	1.2062±0.0652	54.8	0.652	0.4644±0.0035	35.3	0.770
0.4	1.8972±0.0888	64.6	0.620	0.4995±0.0036	46.7	0.696
0.5	2.7363±0.1130	72.1	0.604	0.5271±0.0039	57.7	0.643
0.6	3.6663±0.1345	77.6	0.597	0.5480±0.0036	67.6	0.611
0.7	4.5307±0.1588	81.5	0.594	0.5629±0.0042	75.7	0.595
0.8	5.4145±0.1885	84.4	0.594	0.5737±0.0039	81.7	0.589
0.9	6.2709±0.1994	86.3	0.594	0.5816±0.0036	85.4	0.589
1.0	6.9863±0.2000	87.5	0.595	0.5866±0.0040	87.5	0.591

Table S16: Numerical Values for Fig. S5: Spacegroup $I2_13$, data set (6), $d_{\min} = 1.50$Å

	WIGL -X X			WIGL -X 0		
X	$\langle R1\rangle \pm\sigma(R1)$	$\langle\Delta\Phi\rangle$	$\langle\rho\rangle$	$\langle R1\rangle \pm\sigma(R1)$	$\langle\Delta\Phi\rangle$	$\langle\rho\rangle$
0.0	0.3136±0.0000	0.0	1.000	0.3136±0.0000	0.0	1.000
0.1	0.3749±0.0038	14.7	0.926	0.3432±0.0017	9.5	0.967
0.2	0.5107±0.0134	28.0	0.810	0.3920±0.0030	18.2	0.901
0.3	0.7217±0.0283	39.4	0.724	0.4422±0.0041	26.9	0.826
0.4	1.0134±0.0393	49.5	0.667	0.4861±0.0052	35.4	0.758
0.5	1.3866±0.0527	58.0	0.630	0.5220±0.0056	43.8	0.701
0.6	1.8187±0.0676	65.1	0.608	0.5499±0.0054	51.9	0.657
0.7	2.2550±0.0817	70.8	0.595	0.5710±0.0052	59.6	0.624
0.8	2.7199±0.0986	75.5	0.587	0.5865±0.0055	66.9	0.602
0.9	3.1844±0.1048	79.3	0.583	0.5971±0.0051	72.9	0.588
1.0	3.5849±0.1153	82.1	0.582	0.6044±0.0051	78.1	0.581

Table S17: Numerical Values for Fig. S5: Spacegroup $I2_13$, data set (6), $d_{\min} = 1.90$Å

	WIGL -X X			WIGL -X 0		
X	$\langle R1\rangle \pm\sigma(R1)$	$\langle\Delta\Phi\rangle$	$\langle\rho\rangle$	$\langle R1\rangle \pm\sigma(R1)$	$\langle\Delta\Phi\rangle$	$\langle\rho\rangle$
0.0	0.1929±0.0000	0.0	1.000	0.1929±0.0000	0.0	1.000
0.1	0.2652±0.0023	10.9	0.948	0.2348±0.0015	7.5	0.972
0.2	0.3851±0.0072	21.1	0.862	0.2979±0.0031	14.5	0.922
0.3	0.5438±0.0160	30.7	0.781	0.3616±0.0044	21.2	0.865
0.4	0.7477±0.0258	39.8	0.714	0.4195±0.0050	27.9	0.807
0.5	1.0013±0.0372	48.0	0.664	0.4700±0.0058	34.6	0.754
0.6	1.2993±0.0486	55.3	0.630	0.5122±0.0061	41.2	0.707
0.7	1.6117±0.0571	61.3	0.608	0.5471±0.0062	47.9	0.667
0.8	1.9713±0.0716	66.9	0.593	0.5753±0.0061	54.4	0.634
0.9	2.3481±0.0823	71.5	0.583	0.5969±0.0065	60.6	0.610
1.0	2.6878±0.0861	75.3	0.577	0.6138±0.0061	66.2	0.593

Table S18: Numerical Values for Fig. S5: Spacegroup $I2_13$, data set (6), $d_{\min} = 2.30$Å

	WIGL -X X			WIGL -X 0		
X	$\langle R1\rangle \pm\sigma(R1)$	$\langle\Delta\Phi\rangle$	$\langle\rho\rangle$	$\langle R1\rangle \pm\sigma(R1)$	$\langle\Delta\Phi\rangle$	$\langle\rho\rangle$
0.0	0.0936±0.0000	0.0	1.000	0.0936±0.0000	0.0	1.000
0.1	0.1735±0.0021	8.3	0.965	0.1461±0.0015	5.9	0.979
0.2	0.2790±0.0052	16.1	0.902	0.2138±0.0028	11.5	0.941
0.3	0.4022±0.0105	23.8	0.832	0.2792±0.0040	17.0	0.896
0.4	0.5455±0.0163	31.2	0.769	0.3401±0.0052	22.4	0.847
0.5	0.7185±0.0255	38.4	0.714	0.3941±0.0058	27.8	0.799
0.6	0.9247±0.0332	45.1	0.671	0.4428±0.0066	33.3	0.753
0.7	1.1475±0.0441	51.2	0.639	0.4848±0.0070	38.8	0.712
0.8	1.4160±0.0558	57.2	0.613	0.5201±0.0079	44.3	0.677
0.9	1.7176±0.0663	62.2	0.597	0.5505±0.0072	49.8	0.645
1.0	1.9873±0.0674	66.5	0.584	0.5748±0.0074	55.2	0.619

Table S19: Numerical Values for Fig. S5: Spacegroup $I2_13$, data set (6), $d_{\min} = 2.70$Å

	WIGL -X X			WIGL -X 0		
X	$\langle R1 \rangle \pm \sigma(R1)$	$\langle \Delta\Phi \rangle$	$\langle \rho \rangle$	$\langle R1 \rangle \pm \sigma(R1)$	$\langle \Delta\Phi \rangle$	$\langle \rho \rangle$
0.0	0.0454±0.0000	0.0	1.000	0.0454±0.0000	0.0	1.000
0.1	0.1142±0.0017	6.0	0.976	0.0939±0.0013	4.4	0.985
0.2	0.1974±0.0040	11.8	0.933	0.1517±0.0028	8.7	0.959
0.3	0.2880±0.0072	17.7	0.881	0.2087±0.0038	12.9	0.925
0.4	0.3854±0.0108	23.5	0.826	0.2618±0.0051	17.0	0.888
0.5	0.4919±0.0151	29.3	0.775	0.3119±0.0059	21.3	0.848
0.6	0.6062±0.0198	35.0	0.729	0.3569±0.0067	25.5	0.809
0.7	0.7284±0.0258	40.5	0.688	0.3987±0.0068	30.0	0.770
0.8	0.8755±0.0315	46.2	0.654	0.4355±0.0080	34.4	0.734
0.9	1.0488±0.0397	51.6	0.626	0.4682±0.0079	39.1	0.700
1.0	1.2062±0.0425	56.4	0.605	0.4966±0.0084	43.8	0.668

Table S20: Numerical Values for Fig. S5: Spacegroup $I2_13$, data set (6), $d_{\min} = 3.10$Å

	WIGL -X X			WIGL -X 0		
X	$\langle R1 \rangle \pm \sigma(R1)$	$\langle \Delta\Phi \rangle$	$\langle \rho \rangle$	$\langle R1 \rangle \pm \sigma(R1)$	$\langle \Delta\Phi \rangle$	$\langle \rho \rangle$
0.0	0.0002±0.0000	0.0	1.000	0.0002±0.0000	0.0	1.000
0.1	0.0677±0.0022	4.57	0.984	0.0496±0.0022	3.3	0.991
0.2	0.1364±0.0037	9.03	0.953	0.0983±0.0028	6.6	0.971
0.3	0.2089±0.0060	13.6	0.913	0.1451±0.0038	9.7	0.948
0.4	0.2853±0.0095	18.2	0.869	0.1910±0.0047	12.9	0.920
0.5	0.3643±0.0124	22.9	0.825	0.2346±0.0052	16.2	0.890
0.6	0.4420±0.0148	27.7	0.780	0.2767±0.0065	19.5	0.858
0.7	0.5256±0.0181	32.3	0.740	0.3159±0.0071	22.9	0.826
0.8	0.6330±0.0243	37.7	0.699	0.3537±0.0075	26.5	0.792
0.9	0.7609±0.0300	42.8	0.665	0.3886±0.0084	30.2	0.760
1.0	0.8743±0.0311	47.3	0.638	0.4214±0.0081	34.2	0.727

Table S21: Numerical Values for Fig. S5: Spacegroup $P\bar{1}$, data set (1), $d_{\min} = 0.44$Å

	WIGL -X X			WIGL -X 0		
X	$\langle R1\rangle \pm\sigma(R1)$	$\langle\Delta\Phi\rangle$	$\langle\rho\rangle$	$\langle R1\rangle \pm\sigma(R1)$	$\langle\Delta\Phi\rangle$	$\langle\rho\rangle$
0.0	0.6903±0.0000	0.0	1.000	0.6903±0.0000	0	1.000
0.1	0.8001±0.0267	0.0	0.517	0.7046±0.0016	0	0.664
0.2	1.0202±0.0684	0.0	0.229	0.7204±0.0023	0	0.340
0.3	1.2645±0.0997	0.0	0.085	0.7294±0.0022	0	0.120
0.4	1.5283±0.1254	0.0	0.030	0.7343±0.0022	0	0.048
0.5	1.8062±0.1583	0.0	0.019	0.7370±0.0021	0	0.040
0.6	2.0584±0.1735	0.0	0.015	0.7389±0.0021	0	0.036
0.7	2.3186±0.1963	0.0	0.011	0.7403±0.0021	0	0.031
0.8	2.5402±0.2113	0.0	0.009	0.7411±0.0022	0	0.025
0.9	2.7719±0.2219	0.0	0.007	0.7419±0.0022	0	0.023
1.0	2.9631±0.2387	0.0	0.006	0.7425±0.0021	180	0.021

Table S22: Numerical Values for Fig. S5: Spacegroup $P\bar{1}$, data set (1), $d_{\min} = 0.55$Å

	WIGL -X X			WIGL -X 0		
X	$\langle R1\rangle \pm\sigma(R1)$	$\langle\Delta\Phi\rangle$	$\langle\rho\rangle$	$\langle R1\rangle \pm\sigma(R1)$	$\langle\Delta\Phi\rangle$	$\langle\rho\rangle$
0.0	0.6712±0.0000	0	1.000	0.6712±0.0000	0	1.000
0.1	0.7456±0.0186	0	0.556	0.6902±0.0022	0	0.716
0.2	0.8746±0.0382	0	0.300	0.7132±0.0035	0	0.436
0.3	0.9998±0.0555	0	0.144	0.7271±0.0031	0	0.213
0.4	1.1305±0.0719	0	0.061	0.7336±0.0027	0	0.079
0.5	1.2607±0.0858	-	0.027	0.7367±0.0027	-	0.039
0.6	1.3847±0.1004	0	0.017	0.7390±0.0027	0	0.028
0.7	1.4897±0.1050	-	0.013	0.7405±0.0025	-	0.023
0.8	1.6146±0.1263	-	0.010	0.7414±0.0027	-	0.021
0.9	1.7513±0.1335	-	0.009	0.7420±0.0027	0	0.018
1.0	1.8528±0.1386	0	0.008	0.7426±0.0026	0	0.016

1.6.2 Centrosymmetric Space Group, Data Set (1)

Numerical values presented in Fig. S5 C+D are listed in Tables S21–S26.

1.7 Influence of Parameter Perturbation on Convergence Rate

Numerical Values for Fig. S6, calculated with data set (1), cut to 1.9 Å resolution are listed in Table S27.

2 Crystallization and Data Collection for Data Sets (8) and (6)

2.1 Elastase Data Set (8)

The elastase data set (8) has been used widely for various tutorials on phasing. The data set can be downloaded from http://shelx.uni-ac.gwdg.de/~tg/teaching/anl-ccp4 as part of the tutorial data for

Table S23: Numerical Values for Fig. S5: Spacegroup $P\bar{1}$, data set (1), $d_{\min} = 0.70$Å

	WIGL -X X			WIGL -X 0		
X	$\langle R1\rangle \pm\sigma(R1)$	$\langle\Delta\Phi\rangle$	$\langle\rho\rangle$	$\langle R1\rangle \pm\sigma(R1)$	$\langle\Delta\Phi\rangle$	$\langle\rho\rangle$
0.0	0.6516±0.0000	0	1.000	0.6516±0.0000	0	1.000
0.1	0.7265±0.0141	0	0.634	0.6848±0.0020	0	0.786
0.2	0.8324±0.0320	0	0.398	0.7103±0.0035	0	0.552
0.3	0.9422±0.0475	0	0.236	0.7273±0.0038	0	0.342
0.4	1.0466±0.0614	0	0.126	0.7365±0.0036	0	0.186
0.5	1.1584±0.0727	0	0.063	0.7418±0.0035	0	0.083
0.6	1.2465±0.0846	0	0.035	0.7452±0.0036	0	0.043
0.7	1.3539±0.0912	0	0.020	0.7474±0.0037	0	0.029
0.8	1.4507±0.1076	0	0.016	0.7492±0.0036	0	0.025
0.9	1.5515±0.1096	0	0.014	0.7505±0.0038	0	0.023
1.0	1.6449±0.1201	0	0.011	0.7515±0.0036	0	0.020

Table S24: Numerical Values for Fig. S5: Spacegroup $P\bar{1}$, data set (1), $d_{\min} = 1.20$Å

	WIGL -X X			WIGL -X 0		
X	$\langle R1\rangle \pm\sigma(R1)$	$\langle\Delta\Phi\rangle$	$\langle\rho\rangle$	$\langle R1\rangle \pm\sigma(R1)$	$\langle\Delta\Phi\rangle$	$\langle\rho\rangle$
0.0	0.5383±0.0000	0	1.000	0.5383±0.0000	0	1.000
0.1	0.6193±0.0105	0	0.801	0.5818±0.0045	0	0.888
0.2	0.7577±0.0339	0	0.596	0.6415±0.0088	0	0.737
0.3	0.9222±0.0565	0	0.435	0.6946±0.0105	0	0.584
0.4	1.1017±0.0782	0	0.312	0.7357±0.0117	0	0.449
0.5	1.2972±0.1018	0	0.216	0.7646±0.0125	0	0.329
0.6	1.4977±0.1265	0	0.148	0.7858±0.0122	0	0.228
0.7	1.6814±0.1425	0	0.097	0.8006±0.0128	0	0.151
0.8	1.8791±0.1604	0	0.061	0.8128±0.0123	0	0.087
0.9	2.0388±0.1694	0	0.040	0.8201±0.0130	0	0.053
1.0	2.2330±0.1940	0	0.028	0.8260±0.0123	0	0.038

Table S25: Numerical Values for Fig. S5: Spacegroup $P\bar{1}$, data set (1), $d_{\min} = 1.80$Å

	WIGL -X X			WIGL -X 0		
X	$\langle R1\rangle \pm\sigma(R1)$	$\langle\Delta\Phi\rangle$	$\langle\rho\rangle$	$\langle R1\rangle \pm\sigma(R1)$	$\langle\Delta\Phi\rangle$	$\langle\rho\rangle$
0.0	0.3527±0.0000	0	1.000	0.3527±0.0000	0	1.000
0.1	0.5344±0.0204	0	0.814	0.4567±0.0096	0	0.885
0.2	0.7792±0.0528	0	0.623	0.5850±0.0184	0	0.754
0.3	1.0674±0.0867	0	0.479	0.7038±0.0245	0	0.632
0.4	1.3727±0.1238	0	0.370	0.8026±0.0292	0	0.524
0.5	1.7122±0.1581	0	0.284	0.8877±0.0321	0	0.425
0.6	2.0751±0.1967	0	0.221	0.9574±0.0341	0	0.339
0.7	2.4284±0.2242	0	0.168	1.0102±0.0362	0	0.273
0.8	2.7939±0.2658	-	0.127	1.0560±0.0338	0	0.212
0.9	3.1555±0.2851	0	0.094	1.0934±0.0360	0	0.156
1.0	3.4939±0.3141	-	0.075	1.1228±0.0367	-	0.117

Table S26: Numerical Values for Fig. S5: Spacegroup $P\bar{1}$, data set (1), $d_{\min} = 2.10$Å

	WIGL -X X			WIGL -X 0		
X	$\langle R1\rangle \pm\sigma(R1)$	$\langle\Delta\Phi\rangle$	$\langle\rho\rangle$	$\langle R1\rangle \pm\sigma(R1)$	$\langle\Delta\Phi\rangle$	$\langle\rho\rangle$
0.0	0.1709±0.0000	0	1.000	0.1709±0.0000	0	1.000
0.1	0.3382±0.0125	0	0.858	0.2874±0.0096	0	0.901
0.2	0.5351±0.0304	0	0.713	0.4214±0.0190	0	0.793
0.3	0.7483±0.0554	0	0.577	0.5432±0.0258	0	0.693
0.4	0.9730±0.0779	0	0.465	0.6543±0.0308	0	0.598
0.5	1.2285±0.1116	0	0.370	0.7486±0.0346	0	0.512
0.6	1.4831±0.1257	0	0.294	0.8252±0.0388	0	0.433
0.7	1.7456±0.1630	0	0.232	0.8900±0.0399	0	0.359
0.8	2.0325±0.1919	0	0.186	0.9431±0.0413	0	0.294
0.9	2.3351±0.2258	-	0.144	0.9851±0.0417	0	0.239
1.0	2.6083±0.2385	-	0.114	1.0163±0.0413	0	0.193

Figure S6: R_{complete} *vs.* number of refinement cycles at various amplitudes of perturbation amplitude converge to the same value, the real R_{complete}. Tested with data set (6'). The number of refinement cycles actually performed is shown by tics along the abscissa.

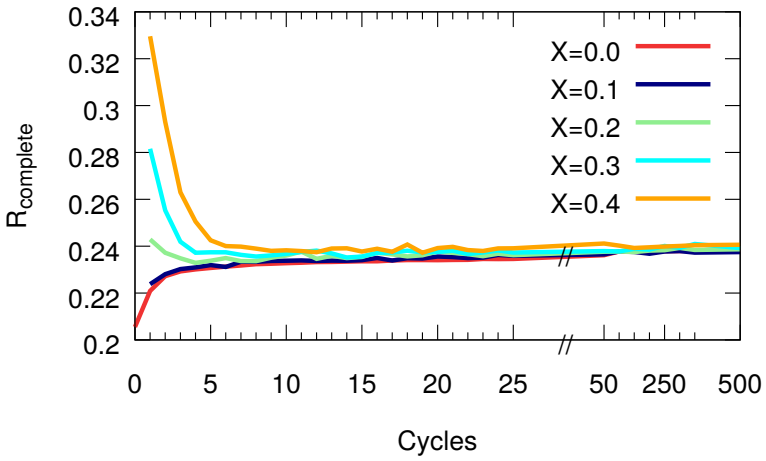


Table S27: Numerical Values for Fig. S6: Value of R_{complete} vs. number of refinement cycles depending on the perturbation amplitude X. 4,000 and 10,000 cycles were only calculated for $X = 0.0$ and $X = 0.3$ to prove convergence within statistical or numerical precision.

#	Perturbation Amplitude X				
	0.0	0.1	0.2	0.3	0.4
0	0.205555				
1	0.220894	0.223783	0.242903	0.281555	0.329551
2	0.227219	0.228039	0.237218	0.255288	0.293071
3	0.229256	0.230216	0.234834	0.241878	0.262984
4	0.230084	0.230963	0.232871	0.237202	0.250519
5	0.230714	0.231944	0.233809	0.237414	0.242529
6	0.231276	0.231155	0.234932	0.237394	0.240037
7	0.231768	0.233354	0.233657	0.236302	0.239823
8	0.232334	0.23334	0.233584	0.235603	0.238969
9	0.23249	0.233718	0.235089	0.236057	0.23801
10	0.232753	0.233813	0.236041	0.236719	0.238237
11	0.232942	0.233918	0.237673	0.237529	0.237896
12	0.233214	0.233695	0.23461	0.238121	0.23742
13	0.233275	0.233853	0.235896	0.236941	0.239012
14	0.233425	0.233811	0.23493	0.235174	0.239078
15	0.233548	0.234037	0.23539	0.235585	0.237742
16	0.233576	0.234985	0.237207	0.237118	0.238928
17	0.23388	0.233906	0.236475	0.237314	0.237628
18	0.234073	0.234874	0.235656	0.23804	0.240712
19	0.234065	0.234717	0.236112	0.23734	0.237193
20	0.234005	0.235592	0.237739	0.237524	0.239199
21	0.234106	0.235376	0.23717	0.238034	0.239732
22	0.234186	0.234913	0.236598	0.236663	0.238334
23	0.234535	0.23511	0.235754	0.237303	0.237951
24	0.234522	0.236237	0.23678	0.238147	0.239057
25	0.234571	0.235888	0.236152	0.237297	0.239077
50	0.236163	0.236475	0.237622	0.237922	0.241152
100	0.237724	0.23776	0.237708	0.237754	0.240228
150	0.237774	0.23753	0.237493	0.238846	0.239245
200	0.237673	0.236798	0.238123	0.238546	0.239541
250	0.237882	0.237746	0.238299	0.240072	0.239877
300	0.237911	0.23792	0.239234	0.239266	0.240103
350	0.238038	0.23725	0.238411	0.24094	0.240429
500	0.238062	0.237508	0.238673	0.239202	0.240602
4,000	0.238284	-	-	0.239047	-
10,000	0.238147	-	-	0.238533	-

Table S28: Data statistics for data set (8).

Space group	$P2_12_12_1$
Cell	a=49.704Å, b=57.895Å, c=74.169Å, $\alpha = \beta = \gamma 90^\circ$
Resolution (outer shell)	37.71 Å–1.37 Å (1.47 Å–1.37 Å)
# refl.	44784 (7662)
Completeness	97.4% (87.2%)
$I/\sigma(I)$	45.16 (7.87)

Table S29: Data statistics for insulin data set (6). Cell: $a = b = c = 77.701$Å$, \alpha = \beta = \gamma = 90^\circ$ Numbers in brackets refer to highest resolution shell.

SG	Cell	Dist. [mm]	$\Delta\phi$	ϕ_{total}	Resol. [Å]	# refl.	multipl.	R_{meas}	$\langle I/\sigma(I)\rangle$	CC(1/2)
$I2_13$	77.701Å	170.28	0.1°	180°	32.13–1.10	32,977	19.81	4.00%	26.2	1.00
					(1.13 – 1.10)	(2,405)	(19.46)	(170.1%)	(1.73)	(0.62)

the CCP4 Workshop Chicago.

Porcine Elastase was purchased from Boehringer Mannheim. The powder was dissolved in 20mM HEPES pH 8.0 and 50–400μl Na_2SO_4 to a nominal protein concentration of 40 mg/ml. Crystals were grown by the hanging drop vapor diffusion. The data were collected with an inhouse Cu rotating anode.

Data statistics shown in Table S28. Data only available as merged data, hence scaling and merging statistics are not available.

2.2 Insulin Data Set (6)

Lyophilized insulin from bovine pancreas was ordered from Sigma–Aldrich. The protein powder was dissolved in 20mM Na_2PO_4 pH=10.0, 10mM Na_3EDTA to a nominal concentration of $20g/l$. Crystals were grown with the sitting drop vapor diffusion method with $2\mu l$ protein solution and $2\mu l$ buffer reservoir. The buffer reservoir contained 275mM Na_2PO_4 pH=10.0 and 10mM Na_3EDTA. Crystals grew within two days. For macro–seeding a selected crystal was transferred into a freshly prepared drop of identical conditions. The final crystal dimensions were $0.23 \times 0.45 \times 0.45mm^3$. For cryo protection the crystal was transferred into the buffer reservoir solution containing 20 % glycerol for three minutes and flash–cooled in liquid nitrogen. Data were collected at BESSY beamline MX 14.1 with a Pilatus 6M detector from Dectris at wavelength $\lambda = 0.826568$ Å. Data were processed with XDS [8]. Details shown in Table S29.

3 Scripts

3.1 Regular Grid into Asymmetric Unit

The following script fills the asymmetric unit of a crystal in space group $I2_13$ with identical atoms suitable for a `shelxl` ins-file. The positions are equally spaced with respect to fractional coordinates, not with respect to Euclidean distance.

```
ACNT=9
RESI=1

printf "%4s%5d%6s\n" "RESI" ${RESI} "UNK"
for x in $(seq 0 99); do
    xc=$(echo "scale=5; 0.5*$x/100" | bc)
    for y in $(seq 0 99); do
        yc=$(echo "scale=5; 0.5*$y/100" | bc)
        for z in $(seq 0 99); do
          if  [[ $z -gt $x  ||  $z -gt $y ]] ; then
            continue;
          fi
          zc=$(echo "scale=5; 0.5*$z/100." | bc)
          if [ $ACNT -gt 99 ]; then
            RESI=$(($RESI+1))
            ACNT=0
            printf "%4s%5d%6s\n" "RESI" ${RESI} "UNK"
          fi
          ACNT=$(($ACNT+1))
          echo "C$ACNT  2   $xc $yc $zc  1.00 0.2"
        done
    done
done
```

The next script was used for the centrosymmetric case of data set (1):

```
#!/bin/bash
# Script to fill a P-1 asymmetric unit with (roughly)
# equal distance atoms
# acnt: atom counter; resi: residue counter

ACNT=9
RESI=1

#P-1 : 0 <=x <= 0.5, 0 <=y <=1; 0 <= z < 1
# Heli data set: C104 P4 S4 H76
# 56 atoms for parameters in unit cell
NUMX=3
NUMY=4
NUMZ=5

printf "%4s%5d%6s\n" "RESI" ${RESI} "UNK"
for x in $(seq 1 $NUMX); do
    xc=$(echo "scale=5; 0.5*$x/$NUMX" | bc)
    for y in $(seq 1 $NUMY); do
        yc=$(echo "scale=5; 1.0*$y/$NUMY" | bc)
        for z in $(seq 1 $NUMZ); do
          zc=$(echo "scale=5; 0.5*$z/$NUMZ" | bc)
          if [ $ACNT -gt 99 ]; then
```

```
            RESI=$(($RESI+1))
            ACNT=0
            printf "%4s%5d%6s\n" "RESI" ${RESI} "UNK"
          fi
          ACNT=$(($ACNT+1))
          printf "C%03d 1 %6.4f %6.4f %6.4f 11.00 0.2\n" \
          $ACNT $xc $yc $zc
        done
    done
done
```

3.2 $R_{complete}$ based Electron Density Maps

The program Coot [2] reads an fcf-file from SHELXL produced with the LIST 6 command. As of version 2014/8, SHELXL provides the LIST 9 command that writes data required for map calculation only for the free set of reflections. LIST 6 contains the data items $hklF_o^2\sigma(F_o^2)F_c\phi_c$ per line. LIST 9 writes the data items $hklF_o^2\sigma(F_o^2)F_c^2\phi_c d$ and the header ends with the term " _shelx_refinement_sigma". The conversion from the former to the latter is achieved with the following script:

```
Fmax=$(grep  -m1 _shelx_F_calc_maximum kcross_set*fcf| \
  sort -g -k2 | tail -n1 | awk '{print $2}')

echo "Fcalc and Fobx will be scaled to Fmax = $Fmax"
Fcmax=$(printf "_shelx_F_calc_maximum%12.2f\n" ${Fmax})

rm -f freemap.fcf
# Create header for fcf file
echo "#" > freemap.fcf
echo "# freemap.fcf created from Rcomplete calculation with LIST 9" >>freemap.fcf
echo "#" >> freemap.fcf
sed -n "1,/^ _refln_F_squared_sigma/p" kcross_set0000.fcf >> freemap.fcf
sed -i "s/^_shelx_refln_list_code      9/_shelx_refln_list_code     6/" freemap.fcf
sed -i "s/^_shelx_F_calc_maximum.*$/${Fcmax}/" freemap.fcf
echo " _refln_F_calc" >> freemap.fcf
echo " _refln_phase_calc" >> freemap.fcf

for i in kcross_set*.fcf; do
  F=$(grep  -m1 _shelx_F_calc_maximum $i | awk '{print $2}')
  s=$(echo "scale=8; $Fmax / $F" | bc)
  sed "1,/^ _shelx_refinement_sigma/d" $i | \
  awk -v S=$s '{print $1, $2, $3, $4, $5, S*sqrt($6), $7}' >> freemap.fcf
done
```

3.3 Data Statistics from Overfitted Data

The following script was used to run shelxl 500 times in order to calculate the data for Tables S15–S26:

```
#!/bin/bash

NUMRUNS=500
INSFILE=$1

rm -f ${INSFILE}.phidata
```

```
rm -f ${INSFILE}.r1data
touch ${INSFILE}.phidata

for i in $(seq 1 ${NUMRUNS}); do
        shelxl  $INSFILE && grep -m1 "^ R1 =" ${INSFILE}.lst \
         >>  ${INSFILE}.r1data
        sed '1,/ _shelx_refinement_sigma/d'${INSFILE}.fcf| \
         awk '{ print $7}' > temp.phidata
        paste temp.phidata ${INSFILE}.phidata > temp2.phidata
        mv temp2.phidata ${INSFILE}.phidata
done

rm temp.phidata
```

Average $R1$ and standard deviation were calculated from the resulting file with the suffix "`r1data`" with the script

```
#!/bin/bash

INSFILE=$1

if [ ! -f "${INSFILE}.r1data" ]
        then
        echo "File ${INSFILE}.r1data does not exist"
        exit
fi

cut -c47-52 ${INSFILE}.r1data > myr1.data
/bin/echo -n "$INSFILE: "
octave --no-window-system -q << eof
r1 = load ('myr1.data');
AveR1 = statistics(r1);
printf ("R1 = %6.4f +/- %6.4f, Data points: %d\n", \
  AveR1(6), AveR1(7), length(r1))
eof
```

The angular statistics were calculated from the file with the suffix "`phidata`" with the following R-script:

```
library(circular, quietly=TRUE, verbose=FALSE)
options (warn=-1)
args <- commandArgs(trailingOnly = TRUE)

ref <- read.table ("wigl00.phidata")
ref <- circular(ref, type="angles", units="degrees", modulo="2pi")

wigl <- read.table (args[1])
wigl <- circular(wigl, type="angles", units="degrees", modulo="2pi")

means  <- vector()
means <- circular(means, type="angles", units="degrees", modulo="2pi")

rhos <- vector()
cts180 <- vector()

for (c in 1:ncol(wigl)) {
```

```
        deltaPhi <- pmax(ref, wigl[,c]) - pmin (wigl[,c] , ref)
        deltaPhi <- pmin (deltaPhi, 360-deltaPhi)
        # count the number of differences for centrosymmetric data
        # should be close to 0 for random data
        cts180 <- c(cts180, sum (deltaPhi==180) - sum(deltaPhi==0))
        means <- c.circular (means, mean.circular(deltaPhi))
        rhos <- c (rhos, rho.circular(deltaPhi))
}

meanphi <- mean.circular (means)
meanrho <- mean (rhos)
print (c(args[1], "Angular mean =", round(meanphi,digits=1)))
print (c(args[1], "Angular rho  =", round(mean(rhos), digits=3)))
```

4 Coordinate Error and Outlier Detection

Full–matrix least squares refinement as implemented in SHELXL allows the calculation of parameter errors. This is routinely used for small molecule structures but also for macromolecular studies, e.g. [9, 10, 11]. The memory requirement of this method increases quadratically with the number of parameters and thus is often not feasible for large crystallographic structures. Errors are instead often estimated based on the Luzzati statistics [12]. Like Wilson, Luzzati only took coordinate errors into account, but not errors in the atomic displacement parameters. The limits of this assumption have been presented above. Several further developments exist based on statistical assumptions [13, 14, 15]

The large number of coordinate files that are created during the calculation of R_{complete} allow the calculation of coordinate errors without statistical assumptions.

The program `CrossCheck` by JL analyses all coordinate files, creates a new file in PDB-format with average coordinates and omitting all 3σ outliers, and lists the atoms with large standard deviations. Table S30 shows an example from the 1.9 Å insulin structure. The list of outliers is very similar. However, with perturbing the parameters much greater deviations even beyond the original amplitude are present and thus more easily detected. In this example, several water molecules could be detected that are only added to noise peaks, and the side chain of lysine B29, residue 2029 in the `SHELXL` ins-file, should clearly be omitted from the model. Such a list has great value for very large models where poorly modelled side chains may be easily overlooked.

References

[1] Sheldrick, G. (2008) A short history of *SHELX*. *Acta Crystallogr.* **A**64, 112–122.

[2] Emsley, P, Lohkamp, B, Scott, W, & Cowtan, K. (2010) Features and development of *Coot*. *Acta Crystallogr.* **D**66, 486–501.

[3] Tickle, I, Laskowski, R, & Moss, D. (2000) Rfree and the Rfree ratio. II. Calculation of the expected values and variances of cross-validation statistics in macromolecular least-squares refinement. *Acta Crystallogr.* **D**56, 442–450.

[4] Sheldrick, G. (2010) Experimental phasing with *SHELXC/D/E*: combining chain tracing with density modification. *Acta Crystallogr.* **D**66, 479–485.

Table S30: Example Output from the program `CrossCheck`. Refinement without (`WIGL 0 0`) and with (`WIGL -0.4 -0.4`) perturbation of the parameters. Perturbation appears to result in larger fluctuations making it easier to identify too optimistically modelled atoms.

`WIGL 0 0`				`WIGL -0.4 -0.4`			
Type	Resi	#	ESD [Å]	Type	Resi	#	ESD [Å]
O	HOH	23	0.6	O	HOH	33	4.6
O	HOH	26	0.3	O	HOH	26	0.6
O4 A	PO4	101	0.2	O	HOH	23	0.6
NZ	LYS	2029	0.1	NZ	LYS	2029	0.5
O3 A	PO4	101	0.1	CD	LYS	2029	0.4
CE	LYS	2029	0.1	O	HOH	28	0.4
O1 A	PO4	101	0.1	O	HOH	4	0.4
O2 A	PO4	101	0.1	O	HOH	21	0.4

[5] Adams, P, Afonine, P, Bunkóczi, G, Chen, V, Davis, I, Echols, N, Headd, J, Hung, L.-W, Kapral, G, Grosse-Kunstleve, R, MCoy, A, Moriarty, N, Oeffner, R, Read, R, Richardson, D, Richardson, J, Terwilliger, T, & Zwart, P. (2010) *PHENIX*: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr* **D**66, 213–221.

[6] Kleywegt, G. (2002) *Uppsala Software Factory* (Uppsala University, Dept. of Cell. and Mol. Biology).

[7] Winn, M, Ballard, C, Cowtan, K, Dodson, E, Emsley, P, Evans, P, Keegan, R, Krissinel, E, Leslie, A, McCoy, A, McNicholas, S, Murshudov, G, Pannu, N, Potterton, E, Powell, H, Read, R, Vagin, A, & Wilson, K. (2011) Overview of the *CCP*4 suite and current developments. *Acta Crystallogr* **D**67, 235–242.

[8] Kabsch, W. (2010) Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr.* **D**66, 133–144.

[9] Köpfer, D, Sing, C, Gruene, T, Sheldrick, G, Zachariae, U, & de Groot, B. (2014) Ion permeation in K^+ channels occurs by direct coulomb knock-on. *Science* **3**46, 352–355.

[10] Gruene, T, Hahn, H, Meilleur, F, & Sheldrick, G. (2014) Refinement of macromolecular structures against neutron data with shelxl-2013. *J Appl Crystallogr* **4**7, 462–466.

[11] Meyer, D, Neumann, P, Koers, E, Sjuts, H, LÃijdtke, S, Sheldrick, G. M, Ficner, R, & Tittmann, K. (2012) Unexpected tautomeric equilibria of the carbanion-enamine intermediate in pyruvate oxidase highlight unrecognized chemical versatility of the thiamin cofactor. *Proc Natl Acad Sci U S A* **1**09, 10867–10872.

[12] Luzzati, P. (1952) Traitement statistique des erreurs dans la determination des structures cristallines. *Acta Crystallogr.* **5**, 802–810.

[13] Cruickshank, D. (1999) Remarks about protein structure precision. *Acta Crystallogr* **D**55, 583–601.

[14] Read, R. (1986) Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr.* **A**42, 140–149.

[15] Gurusaran, M, Shankar, M, Nagarajan, R, Helliwell, J, & Sekar, K. (2014) Do we see what we should see? Describing non-covalent interactions in protein structures including precision. *IUCrJ* **1**, 74–81.