Revisiting the "satisfaction of spatial restraints" approach of MODELLER for protein homology modeling

Giacomo Janson[1*], Alessandro Grottesi[2], Marco Pietrosanto[3], Gabriele Ausiello[3], Giulia Guarguaglini[4¶], Alessandro Paiardini[1*¶]

[1]Department of Biochemical Sciences "A. Rossi Fanelli", "Sapienza" University of Rome, Roma, Italy

[2]Super Computing Applications and Innovation (CINECA), Roma, Italy

[3]Centre for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata, Roma, Italy

[4]Department of Biology and Biotechnology, Institute of Molecular Biology and Pathology, "Sapienza" University of Rome, Roma, Italy

*Corresponding authors

[¶]These authors contributed equally to this work.

E-mail: giacomo.janson@uniroma1.it, alessandro.paiardini@uniroma1.it

Abstract

The most frequently used approach for protein structure prediction is currently homology modeling. The 3D model building phase of this methodology is critical for obtaining an accurate and biologically useful prediction. The most widely employed tool to perform this task is MODELLER. This program implements the "modeling by satisfaction of spatial restraints" strategy and its core algorithm has not been altered significantly since the early 1990s. In this work, we have explored the idea of modifying MODELLER with two effective, yet computationally light strategies to improve its 3D modeling performance. Firstly, we have investigated how the level of accuracy in the estimation of structural variability between a target protein and its templates in the form of σ values profoundly influences 3D modeling. We show that the σ values produced by MODELLER are on average weakly correlated to the true level of structural divergence between target-template pairs and that increasing this correlation greatly improves the program's predictions, especially in multiple-template modeling. Secondly, we have inquired into how the incorporation of statistical potential terms (such as the DOPE potential) in the MODELLER's objective function impacts positively 3D modeling quality by providing a small but consistent improvement in metrics such as GDT-HA and lDDT and a large increase in stereochemical quality. Python modules to harness this second strategy are freely available at https://github.com/pymodproject/altmod. In summary, we show that there is a large room for improving MODELLER in terms of 3D modeling quality and we propose strategies that could be pursued in order to further increase its performance.

Author summary

Proteins are fundamental biological molecules that carry out countless activities in living beings. Since the function of proteins is dictated by their three-dimensional atomic structures, acquiring structural details of proteins provides deep insights into their function. Currently, the most successful computational approach for protein structure prediction is template-based modeling. In this approach, a

target protein is modeled using the experimentally-derived structural information of a template protein assumed to have a similar structure to the target. MODELLER is the most frequently used program for template-based 3D model building. Despite its success, its predictions are not always accurate enough to be useful in Biomedical Research. Here, we show that it is possible to greatly increase the performance of MODELLER by modifying two aspects of its algorithm. First, we demonstrate that providing the program with accurate estimations of local target-template structural divergence greatly increases the quality of its predictions. Additionally, we show that modifying MODELLER's scoring function with statistical potential energetic terms also helps to improve modeling quality. This work will be useful in future research, since it reports practical strategies to improve the performance of this core tool in Structural Bioinformatics.

Introduction

In silico protein structure prediction constitutes an invaluable tool in Biomedical Research, since it allows to obtain structural information on a large number of proteins currently lacking an experimentally-determined 3D structure [1]. Template-based modeling (TBM) has been shown to be the most practically useful prediction strategy [2].

Homology modeling (HM) is a fast and reliable TBM method in which a target protein is modeled by using as a structural template an homologous protein. HM predictions usually consist of three phases. In the first, the sequence of the target is used to search for suitable templates in the PDB [3-4]. In the second, a sequence alignment between the target and templates is built with the goal of inferring the equivalences between their residues [5]. In the final, the information of the templates is used to build a 3D atomic model of the target.

The overall accuracy of HM has remarkably increased in the last 25 years [6]. This has been promoted mostly by advances in template searching and alignment building algorithms, while only minor

advances have been witnessed in the 3D model building step [7]. However, recent breakthroughs in protein structure refinement methods [8-9] envisage a large room for improvement in HM which could originate from advances in 3D model building.

MODELLER [10] is the most frequently used program for 3D model building in HM. One of the main reasons of its success has been its accurate [11], yet fast algorithm. In MODELLER, the information contained in an input target-template alignment is used to generate a series of homology-derived spatial restraints (HDSRs), acting on the atoms of the 3D protein model. Sigma ("σ") values of homology-derived distance restraints (HDDRs) determine the amount of conformational freedom which the model is allowed to have with respect to its templates. MODELLER uses a statistical "histogram-based" strategy to estimate σ values [12]. These restraints are incorporated into an objective function which also includes physical energetic terms from CHARMM22 [13]. A fast, but effective optimization algorithm based on a combination of conjugate gradients (CG) and molecular dynamics with simulated annealing (MDSA) is then used to identify a model conformation that satisfies as much as possible the HDSRs, while retaining stereochemical realism.

The core MODELLER algorithm was developed in the early 1990s and it was essentially left unchanged over the years. Despite its importance, there have been relatively few attempts to improve it.

In 2015, Meier and Söding designed a novel probabilistic framework for building HDDRs [7], whose aim was to help MODELLER tolerate alignment errors and to combine the information from multiple templates in a statistically rigorous way. This system increased 3D modeling quality, especially for multiple-template modeling. However, since it is integrated in the HHsuite project [14] it can be employed only when the first two phases of HM are carried out by programs of the HHsuite package.

Researchers from Lee's group developed a modified version of MODELLER which they have been using in CASP experiments [15-17]. First, they replaced the MODELLER optimization algorithm with

the more thorough conformational space annealing (CSA) method [18]. Secondly, they pioneered a new strategy to assign σ values to HDDRs relying on machine learning [19]. Finally, they included a series of additional terms to the MODELLER objective function, such as terms for the DFIRE [20] and DFA [21] knowledge-based potentials, for hydrogen bond formation [22] and to enforce in models predictions of structural properties. In terms of 3D modeling quality, this system outperformed the original MODELLER [17]. Unfortunately, the separated contribution of several of these modifications is not reported and much of this system remains in-house (only the CSA algorithm is publicly available).

Although these seminal studies have shown that the core MODELLER algorithm has room for improvement, most of its users employ its original version, probably because existing modifications either depend on additional packages to install, or are computationally too expensive (e.g., the CSA algorithm alone was reported to increase computational times by a factor of ~130). Since MODELLER is a core tool in Structural Bioinformatics, it is of paramount importance to investigate in detail the inner working of its algorithm and to develop it further. Here, we have explored two computationally light strategies to improve it in terms of 3D modeling quality.

Particular attention has been dedicated in understanding how the level of accuracy in the estimation of structural variability between the target and templates expressed as σ values influences 3D modeling. Although in this work we have not modified the MODELLER algorithm for σ values assignment, we propose strategies that could be likely pursued in the next-future in order to greatly increase the performance of the program. Additionally, we have investigated how the incorporation of statistical potential terms, such as DOPE [23], in the program's objective function is able to impact positively 3D modeling and under certain conditions (for example in single-template modeling) it can be coupled synergistically to the previous strategy.

To rigorously validate these approaches, we have benchmarked them using protein targets from a diverse set of high-resolution structures from the PDB and we quantified the individual impact on 3D modeling of each modification. This information will be useful in future research, since it shows in which areas there is still room for improvement and in which areas it might be difficult to advance further.

Materials and methods

Outline of MODELLER's homology-derived distance restraints

The MODELLER approach relies on the generation of HDSRs for interatomic distances and dihedral angles [12]. Each HDSR is treated as a probability density function (*pdf*). HDSRs acting on interatomic distances (that is, HDDRs) have a predominant role in determining the 3D structure of a model. The way they are built is summarized here.

For a couple of atoms *i* and *j* of the model, the program finds in the template the equivalent atoms *k* and *l* which have a distance in space of d_t. The distance d_m between *i* and *j* is assumed to be normally distributed around d_t with a standard deviation σ and the *pdf* restraining it is:

$$f(d_m) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(d_m - d_t)^2}{2\sigma^2}} \quad . \qquad (1)$$

In MODELLER *pdfs* are converted in objective function terms as follows:

$$obj(d_m) = -\ln\left(f(d_m)\right) = -\ln\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(d_m - d_t)^2}{2\sigma^2}}\right) = \frac{(d_m - d_t)^2}{2\sigma^2} - \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) \quad , \qquad (2)$$

therefore Gaussian HDDRs correspond to harmonic potential terms. Since HDDRs are considered to be independent, their objective function terms are summed. HDDRs are built for four groups of atoms: the Cα-Cα, backbone NO, side chain-main chain (SCMC) and side chain-side chain (SCSC) groups (see **S1**

Table). MODELLER generates its σ values (hereinafter named σ_{MOD} values) through an histogram-based approach [12].

MODELLER allows to take advantage of multiple templates, a strategy that (when templates are chosen adequately) usually outperforms single-template modeling [24]. When employing U templates to restrain a distance d_m, MODELLER uses the following *pdf*:

$$f(d_m) = \sum_{u=1}^{U} w_u \frac{1}{\sigma_u \sqrt{2\pi}} e^{-\frac{(d_m - d_{t,u})^2}{2\sigma_u^2}} , \quad (3)$$

where u is the template index, w_u is a template-specific weight, $d_{t,u}$ and σ_u are the distance observed in template u and its σ value respectively. In MODELLER, w_u is a function of the local sequence similarity between the target and template u.

The total objective function of MODELLER (F_{TOT}) can be expressed as follows:

$$F_{TOT} = F_{PHYS} + F_{HOM} , \quad (4)$$

where F_{PHYS} contains five physical terms (see **S2 Table**) and F_{HOM} contains HDSRs terms. In this work, the weights for F_{PHYS} and F_{HOM} were always left to 1.0 (therefore they are omitted from the formula above).

Benchmarking MODELLER modifications with an analysis set

In order to benchmark modifications of MODELLER, we built an analysis set of selected target proteins. We obtained 926 X-ray structure chains from PISCES [25], using the following criteria to filter the PDB:

- the maximum mutual sequence identity (SeqId) among the chains was 10%;
- their structures had a resolution < 2.0 Å and R-factor < 0.25;

- they contained no missing residues due to lacking electron density;
- their length was between 70 and 700 residues.

These chains were our target candidates. To obtain their templates, we culled from PISCES another set using similar filters, except that this time the maximum mutual SeqId was 90%. We removed from this larger set all the targets, obtaining 6224 chains. Each target was then aligned to these chains using TM-align [26] and we selected as template candidates the chains meeting the following criteria:

- the SeqId in the structural alignment built by TM-align was between 15% and 95%;
- the two TM-scores [27] produced by TM-align (each score is normalized by the length of one of the aligned proteins) were at least 0.6, a threshold to consider two proteins as homologous [28].

We retained for each target only its top five templates in terms of TM-score (normalized on the target length). In this way, we obtained a final set of 225 target chains (suitable templates could not be found for 701 targets, a result of using only high-resolution template structures). For each target, we performed single-template modeling only with its top template and therefore we had 225 single-template models, which constituted the Analysis Single-template (AS) set. 118 targets had at least two templates (with an average of 3.3), which constituted the Analysis Multiple-templates (AM) set.

The average SeqId for the AS target-template alignments is 0.38. Improving the performance of MODELLER with targets having templates with a SeqId < 0.40 is important, because these cases are the most frequent ones in Biomedical Research [29] and the accuracy of TBM is often low in this regimen. The well-equilibrated distributions of SeqId, target coverage, target length and of CATH structural classes [30] of the analysis set (see **S1 Fig**) assure that our results have a general validity.

Alignment building

In order to align target-template pairs we employed the accurate HHalign program [4], which confronts two profile hidden Markov models. To build input profiles for HHalign, we ran HHblits [31] with its default parameters and three search iterations against the *uniprot20_2016_02* database. After employing HHalign to align pairs of target-template profiles, we extracted from the program's output their pairwise alignments. Multiple target-templates alignments were obtained by joining pairwise alignments.

Whenever specified, we also employed target-template alignments built with TM-align in order to assess the effect on 3D modeling of HDDRs derived from error-free structural alignments.

3D model building and evaluation

For all benchmarks we used MODELLER version 9.21. In order to modify its objective function terms, restraints parameters and optimization schedules we interfaced with its Python API.

In MODELLER, the final quality of a model is largely determined in the MDSA phase. In this work, unless otherwise stated, we employed the default *very_fast* MDSA protocol of the program (corresponding to a 5.4 ps run). When specified, we also employed the more thorough *slow* protocol (corresponding to a 18.4 ps run). The CG protocol was always left to its default parameters.

The approach used to evaluate the quality of an homology model was to build 16 different copies of it (hereinafter defined as decoys), and to report as an overall quality score (see below) the average score of the 16 decoys.

To evaluate the quality of the backbones we used the GDT-HA metric [6] computed by the TM-score program. In order to evaluate the quality of local structures and side chains, we used the lDDT metric [32], computed by the lDDT program. Detailed descriptions of these two metrics are given in **S1 Text**. To evaluate the stereochemical quality of models we employed MolProbity scores computed by the

MolProbity suite [33]. A MolProbity score expresses the global stereochemical quality of a 3D model. The lower it is, the higher is the quality of the model.

Optimal σ values for homology-derived distance restraints

σ values of HDDRs have a fundamental role in MODELLER. A natural question is: given a target-template alignment, what is the set of σ values which will maximize 3D modeling accuracy? The concept of optimal σ values in single-template modeling was addressed for the first time by the Lee group [19]. They reported that for a Gaussian HDDR acting on a distance d_m between atoms i and j in a 3D model, the optimal σ value is:

$$|\Delta d_n| = |d_n - d_t| \,, \qquad (5)$$

where d_t is the distance between the template atoms equivalent to i and j and d_n is the distance between i and j observed in the experimentally-determined native target structure. We show that the use of $|\Delta d_n|$ values for Gaussian HDDRs is supported by theory, as it can be analytically proven that they maximize the likelihood of obtaining a model in which each restrained d_m is equal to its corresponding d_n (see **S2 Text**).

In the case of multiple-template HDDRs, we demonstrate that the combination of optimal σ values and weights can be found again analytically (see **S3 Text**). In this situation, the optimal σ values are again $|\Delta d_n|$ values. The associated template weighting scheme assigns a weight of 0 to all templates with the exception of the template with the lowest σ, which should have a weight of 1. We termed this scheme as the "only-lowest" (OL) scheme. Note that the OL scheme is an extreme case of the weighting scheme proposed in [34] (see **S3 Text**).

Whenever using $|\Delta d_n|$ values as σ parameters, we had to modify them by setting their minimum value at 0.05 Å. Raw $|\Delta d_n|$ values are extracted directly from pairs of homologous protein structures and they

are often close to 0 Å (see **Fig 1A**). In MODELLER, HDDRs having very small σ values will seldom be satisfied because their quadratic objective function terms will penalize enormously even minimal deviations from templates. In fact, using unmodified $|\Delta d_n|$ values often leads to modeling failures, since the total objective function of models surpasses the allowed limit of MODELLER, stopping the model building process. Setting a lower limit to their value, allows their use in 3D modeling.

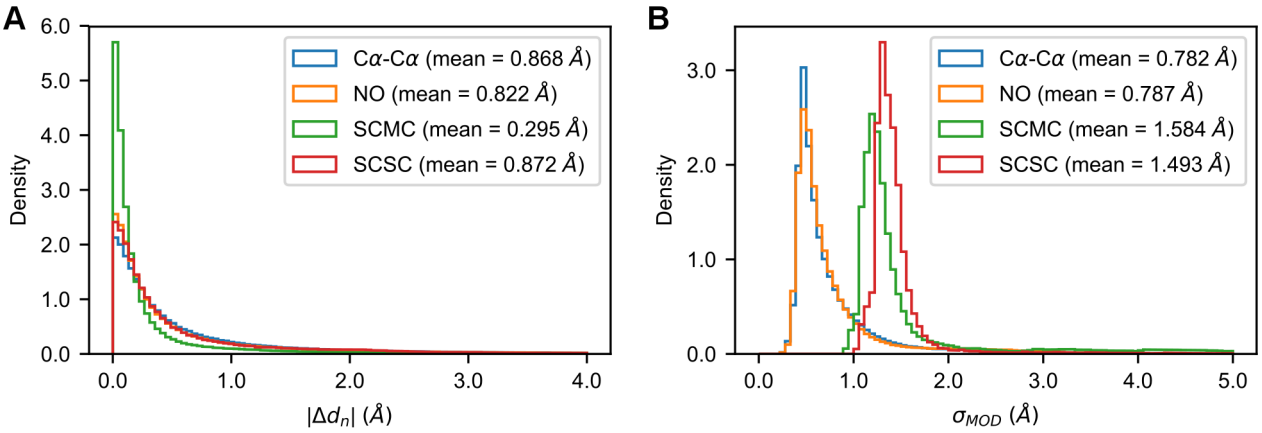


Fig 1. Distribution of $|\Delta d_n|$ and σ_{MOD} values. Distributions of the $|\Delta d_n|$ (A) and σ_{MOD} (B) values observed in the AS models for the four HDDR groups of MODELLER. Beside the names of the restraints groups, their mean values are reported.

Perturbing optimal $|\Delta d_n|$ values

To understand the effect of using error-containing estimations of $|\Delta d_n|$ values on 3D modeling, we randomly selected a fraction f_e of the HDDRs in a target-template pair and substituted their $|\Delta d_n|$ values with randomly generated ones.

Random values were extracted from exponential distributions fitted on the Cα-Cα, NO, SCMC and SCSC $|\Delta d_n|$ data observed in our AS models (see **Fig 1A**). These exponentials well-approximate the

observed $|\Delta d_n|$ distributions and their means were taken to be the same. Since 3D modeling quality tends to decrease when the average σ value of a model increases (see **Fig 2A** and **2B**), this perturbation scheme ensures that when replacing $|\Delta d_n|$ values with random numbers, alterations in the quality of 3D models will not be caused by just changing their mean σ values.

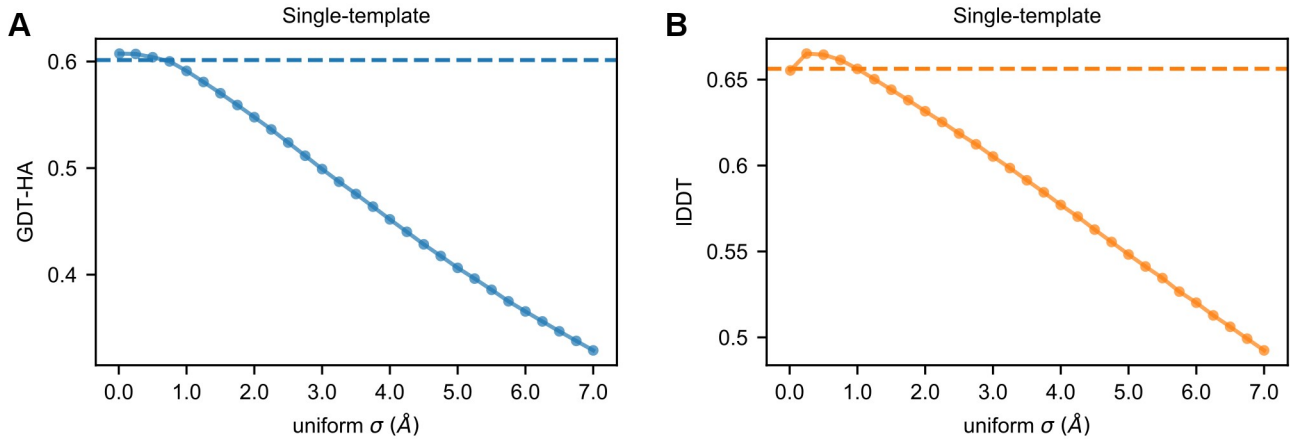


Fig 2. Modeling with uniform σ values. Average GDT-HA (A) and lDDT (B) scores of the AS models as a function of the uniform σ value (ranging from 0.01 to 7.0 Å) applied to their HDDRs. The horizontal dashed lines represent the average scores obtained with the original σ_{MOD} values.

We used 10 f_e values (linearly spacing from 0.1 to 1.0) and for each, we generated 5 sets of perturbed $|\Delta d_n|$ values per target-template pair, which allowed to better sample the effect of perturbations. For each perturbed set, we built 8 decoys (resulting in a total of 5*8=40 decoys for each f_e value). For a certain f_e value, the quality score for a model was recorder as the average score of all its 40 decoys.

To quantify in terms of Pearson correlation coefficient (PCC) the amount of perturbation introduced in the $|\Delta d_n|$ values of a single model, for each f_e we used a score defined as PCC_{MODEL}. This score is computed as:

$$PCC_{MODEL} = \frac{1}{n_R} \sum_{r=1}^{n_R} \left(\frac{1}{U} \sum_{u=1}^{U} PCC\left(d_u, p_{u,r}\right) \right) \quad , \qquad (6)$$

where n_R is the number of perturbed $|\Delta d_n|$ sets (in our case 5), r is the index for these sets, U is the number of templates of the model, PCC indicates the Pearson correlation coefficient, d_u is the list of $|\Delta d_n|$ values associated with the u-th template and $p_{u,r}$ is the list of perturbed $|\Delta d_n|$ values associated with the *u-th* template in set r. For each HDDR group, the relationship between f_e and the average PCC_{MODEL} of the AS and AM sets is roughly linear (**S2 Fig**).

Inclusion of statistical potential terms in the objective function of MODELLER

In this work, we explored the effect of including in the objective function of MODELLER terms for interatomic distance statistical potentials. These potentials are developed with the aim of recognizing native-like protein conformations [35], therefore their use could help MODELLER to approach these conformations [36].

We employed the DOPE potential [23], which is integrated in the MODELLER package where it is commonly used to evaluate qualities of 3D models. DOPE is an “all atom” potential. Its 12561 terms are approximated with interpolating cubic splines, which can be differentiated analytically and used in the gradient-based optimization algorithm of the program.

The Lee group previously included the DFIRE [20] potential in the MODELLER objective function [15]. To compare their performances in 3D model building, we also integrated DFIRE in MODELLER (DFIRE parameters were obtained from its source code).

When including statistical potential terms, the MODELLER objective function becomes:

$$F_{TOT} = F_{PHYS} + F_{HOM} + w_{SP} F_{SP} \quad , \qquad (7)$$

where F_{SP} contains the statistical potentials terms and w_{SP} is their weight. For obtaining best 3D modeling results, we tested several values of w_{SP}.

We employed statistical potentials using a contact shell value of 8.0 Å. Higher values can be safely avoided because the terms of DOPE and DFIRE start to acquire a flat shape over the 8.0 Å threshold (see **S3 FigA**). The code we used to employ these potentials in MODELLER is freely available at https://github.com/pymodproject/altmod.

Results

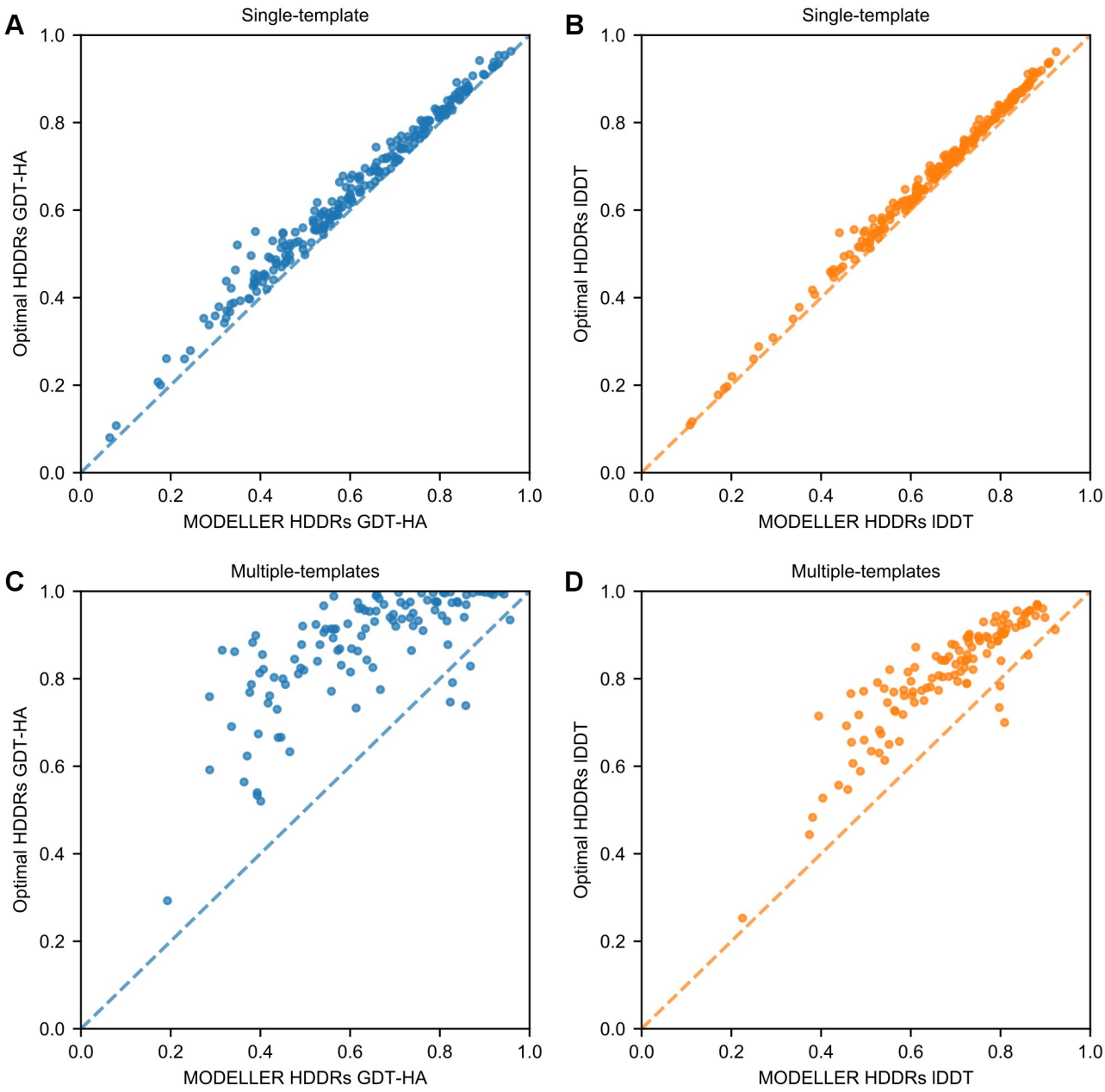
Effects of optimal σ values on 3D modeling

Effects on single-template modeling. Gaussian HDDRs are the heart of the MODELLER approach. At first, we explored how the use of optimal σ values (that is, $|\Delta d_n|$ values) influences single-template modeling. The Lee group already reported it to bring significant improvements for a small number of proteins. Here, we extended the analysis to a larger set to derive general conclusions. As shown in **Table 1**, employing restraints bearing $|\Delta d_n|$ values greatly increases 3D modeling accuracy. In terms of global Cα backbone quality, the average GDT-HA score of the AS models increases by 6.0% with respect to the score obtained with σ_{MOD} values. An improvement is also observed for local all-atom quality, as the average lDDT score increases by 4.2%. Increments in GDT-HA and lDDT are seen for 224/225 and 225/225 AS models respectively (see **Fig 3A** and **3B**).

Table 1. 3D modeling qualities of the AS single-template models built with optimal HDDRs and alignments.

Strategy	GDT-HA	lDDT	MolProbity score
MODELLER[a]	0.6014 (-)	0.6563 (-)	3.0104 (-)
OPTIMAL[b]	0.6377 (+6.0%)*	0.6842 (+4.2%)*	3.0311 (+0.7%)*
MODELLER-SLOW[c]	0.6036 (+0.4%)*	0.6594 (+0.5%)*	2.8512 (-5.3%)*
OPTIMAL-SLOW	0.6377 (+6.0%)*	0.6853 (+4.4%)*	2.9039 (-3.5%)*
MODELLER-TMalign[d]	0.6383 (+6.1%)*	0.6951 (+5.9%)*	3.0411 (+1.0%)*
OPTIMAL-TMalign	0.6805 (+13.2%)*	0.7259 (+10.6%)*	3.0870 (+2.5%)*

The “GDT-HA”, “lDDT” and “MolProbity score” columns report the average values for those metrics. Percent improvements are computed with respect to the scores of the default MODELLER (first row), while asterisks denote a statistically significant difference with respect to them (according to a Wilcoxon signed-rank tests with a significance level of 0.05). See **S3 Table** for a full list of the numerical p-values. [a]The “MODELLER” prefix indicates that the strategy employs HDDRs generated by MODELLER. [b]The “OPTIMAL” prefix indicates the use of optimal HDDRs. [c]The “SLOW” suffix indicates the use of the *slow* MDSA protocol instead of the default *very_fast* one. [d]The “TMalign” prefix indicates the use of target-template alignment built through TM-align.



304 **Fig 3. The use of optimal parameters for HDDRs improves 3D modeling quality.** (A) and (B) GDT-
305 HA and lDDT scores of the AS models built with σ_{MOD} (reported on the x-axis) and with optimal $|\Delta d_n|$
306 (y-axis) values. (C) and (D) GDT-HA and lDDT scores for the AM models obtained with
307 MODELLER-generated (x-axis) and optimal (y-axis) HDDRs.

308

Increasing target-template alignment quality is one of the current challenges in TBM. In our AS models, the average accuracy of HHalign sequence alignments with respect to error-free TM-align structural alignments is 0.87 (see **S4 Fig**). When rebuilding the AS models using σ_{MOD} values and TM-align alignments, the average GDT-HA and lDDT scores improve by 6.1% and 5.9% respectively over the scores obtained with σ_{MOD} values and HHalign alignments (see **Table 1**). These results show that by optimizing parameters of the 3D model building phase of single-template HM, the same improvement obtainable by optimizing alignment building can be reached.

It might be thought that $|\Delta d_n|$ values aid 3D modeling by compensating for alignment errors, that is, by assigning misaligned residues more conformational freedom to help MODELLER repositioning them in a correct way. However, their effect can not be explained only by this mechanism, since they yield a 6.6% and 4.4% increase in GDT-HA and lDDT also when models are built with TM-align alignments (see **Table 1**).

Effects on multiple-template modeling. Next, we explored the effect of optimal HDDRs in multiple-template modeling, which has never been assessed before. As shown in **Table 2**, applying an optimal set of σ values and template weights results in an enormous improvement in the quality of 3D models (see also **Fig 3C** and **3D**). When building the AM models with optimal restraints, their average GDT-HA and lDDT scores improve by 38.9% and 18.9% over the scores obtained by using MODELLER-generated restraints. These increments are larger than the one observed when performing multiple-template modeling with MODELLER-generated restraints and error-free TM-align structural alignments, which result in a 5.7% and 5.1% improvements in GDT-HA and lDDT.

Optimal HDDRs increase even more the beneficial effect of using multiple templates. With MODELLER-generated restraints, employing multiple templates leads to an improvement of 1.9% and 2.0% in the average GDT-HA and lDDT of the AM models over single-template modeling performed

with top-templates (see the MODELLER-ST strategy in **Table 2**). On the other hand, with optimal HDDRs, it leads to an improvement of 33.2% and 16.0% in GDT-HA and lDDT over single-template modeling performed with optimal HDDRs (see the OPTIMAL-ST strategy in **Table 2**).

The reason for this large improvement is the following. In MODELLER, the *pdf* for a multiple-template HDDR includes a weighted contribution from each template. In optimal HDDRs, $|\Delta d_n|$ values are employed as σ values in conjunction with the OL weighting scheme (see the Methods section). In this scheme, only the contribution of the best template is selected for each HDDR (when considering a single HDDR, the best template is defined as the one having a distance d_t as close as possible to the target distance d_n). On the other hand, in MODELLER-generated HDDRs, the weights are usually non-zero for every template, meaning that the contribution of the best template is always weakened. This effect increases the allowed conformational space for the restrained distance, thus making it less likely to build a model with a near-native distance.

The importance of the template-weighting scheme [7] is illustrated by the fact that when employing $|\Delta d_n|$ values and a uniform weighting scheme (that is, for an HDDR with U templates each template is given a weight $w_u = 1/U$), the average GDT-HA and lDDT scores of the AM models improve only by 18.3% and 8.9% over the standard MODELLER (see the OPTIMAL-U strategy in **Table 2**).

Table 2. 3D modeling qualities of the AM multiple-template models built with optimal HDDRs and alignments.

Strategy	GDT-HA	lDDT	MolProbity score
MODELLER	0.6287 (-)	0.6819 (-)	3.0725 (-)
OPTIMAL	0.8733 (+38.9%)*	0.8106 (+18.9%)*	3.1478 (+2.4%)
MODELLER-SLOW	0.6310 (+0.4%)*	0.6850 (+0.5%)*	2.9143 (-5.2%)*
OPTIMAL-SLOW	0.8747 (+39.1%)*	0.8133 (+19.3%)*	3.0475 (-0.8%)*
OPTIMAL-U[a]	0.7438 (+18.3%)*	0.7427 (+8.9%)*	3.1744 (+3.3%)
MODELLER-ST[b]	0.6168 (-1.9%)*	0.6683 (-2.0%)*	3.0231 (-1.6%)*
OPTIMAL-ST	0.6557 (+4.3%)*	0.6986 (+2.5%)*	3.0398 (-1.1%)
MODELLER-TMalign	0.6645 (+5.7%)*	0.7165 (+5.1%)*	3.0529 (-0.6%)
OPTIMAL-TMalign	0.9222 (+46.7%)*	0.8498 (+24.6%)*	3.1044 (+1.0%)

See **Table 1** for the description of contents, columns and most modeling strategies names. See **S4 Table** for a full list of the numerical p-values. [a]The "U" suffix indicates the use of uniform template weights for multiple-template HDDRs. [b]The "ST" suffix indicates that only the top template for each target was used (thus resulting in single-template modeling).

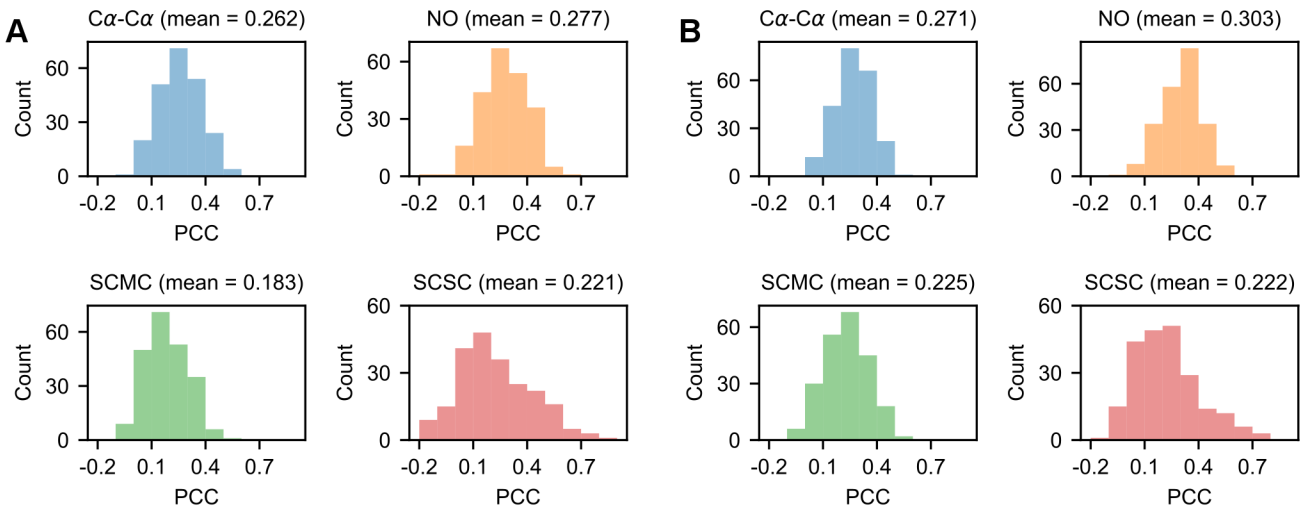
Effects on stereochemical quality. In both single and multiple-template modeling, the use of optimal HDDRs appears to decrease the stereochemical quality of models, as seen by increased MolProbity scores (see **Table 1** and **Table 2**). The increment is more prominent in multiple-template modeling (2.4%) than in single-template modeling (0.7%). While optimal restraints may guide the models in conformations near the native state, at the same time they probably force stereochemical inaccuracies. However, employing a more through MDSA protocol is sufficient to almost entirely relax these

362 inaccuracies, while maintaining high GDT-HA and lDDT scores (see the strategies with the "SLOW"
363 suffix in the tables).

364 **Perturbing optimal σ values**

365 As first demonstrated in [19], σ_{MOD} values are weakly correlated with their optimal counterparts. In the
366 AS models, the distributions of $|\Delta d_n|$ and σ_{MOD} values are markedly different (see **Fig 1A** and **1B**) and
367 the average PCCs between them are 0.262, 0.277, 0.183 and 0.221 for the Cα-Cα, NO, SCMC and
368 SCSC restraints groups respectively (see **Fig 4A**). Even with accurate alignments built through TM-
369 align, the histogram-based approach of MODELLER produces σ values which are weakly correlated to
370 $|\Delta d_n|$ values (see **Fig 4B**).

371



372 **Fig 4. Correlation between σ_{MOD} and $|\Delta d_n|$ values in the AS models.** (A) Distributions for the PCCs
373 between σ_{MOD} and $|\Delta d_n|$ values for the HDDRs of the 225 AS models. (B) PCC distributions for the AS
374 models rebuilt with TM-align alignments.

375

In the previous section we have seen that the use of optimal σ values greatly improves MODELLER's predictions. However, since $|\Delta d_n|$ values can not be directly inferred without the prior knowledge of the actual 3D structure that we are trying to predict, a strategy to improve MODELLER would consist in accurately estimating them. Irrespective of the predictive algorithm, it is reasonable to suppose that $|\Delta d_n|$ estimations will always bear a certain amount of error. In order to understand how 3D modeling quality changes as a function of this error, we rebuilt the models of the analysis set by perturbing their $|\Delta d_n|$ values with random noise.

Effects on single-template modeling. Fig 5A shows how the average GDT-HA of the AS models changes when increasing the fraction of $|\Delta d_n|$ values substituted with a random σ (see **Fig 5B** for the relationship with lDDT). In the absence of any perturbation, the average GDT-HA is at its maximum of 0.6377. When substituting just 10% of the $|\Delta d_n|$ values, the mean Cα-Cα PCC_{MODEL} of the models becomes 0.85 and the average GDT-HA decreases by 2.6%. Further increasing the fraction of random σ values leads to a continuous decrease in quality. When all the restraints have a random σ, the average Cα-Cα PCC_{MODEL} approximates 0 and the average GDT-HA is 0.6052 (resulting in a 5.1% decrease with respect to the optimal state). This score is 0.7% higher than the average GDT-HA obtained using the default σ_{MOD} values, which is 0.6009. Although the difference between these two scores is statistically significant (Wilcoxon signed-rank test, p-value = 3.7e-4) it is only minimal from a structural point of view. In other words, in single-template modeling, provided that the average σ of a model does not surpass a certain threshold (that is, the average $|\Delta d_n|$ observed in nature), randomly generated σ values are surprisingly as effective as those generated by the MODELLER histogram-based approach. This is also confirmed by the fact that the use of uniform σ values < 1.0 Å does not significantly alter the GDT-HA and lDDT scores of models with respect to the standard MODELLER algorithm (see **Fig 2A** and **2B**).

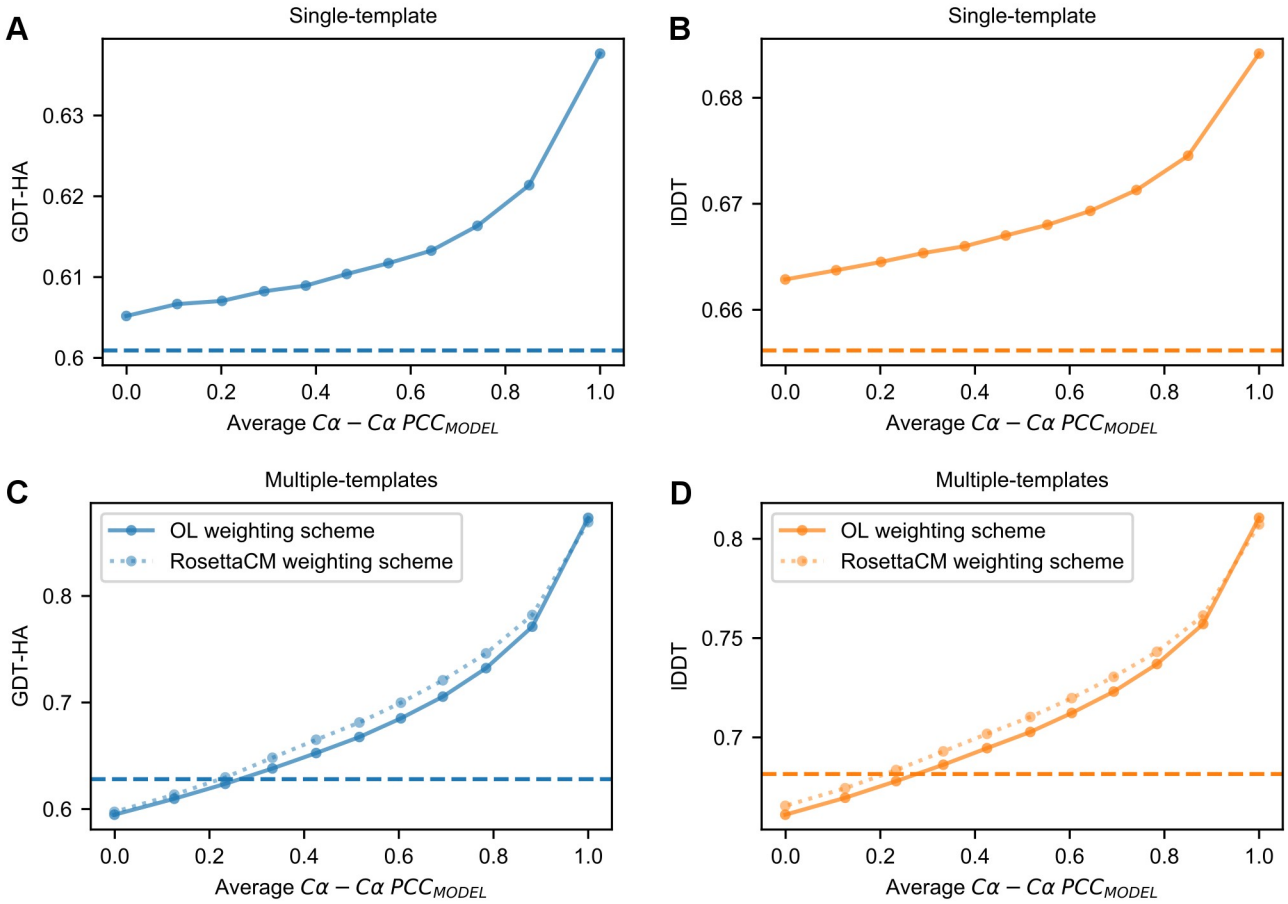


Fig 5. Effect of $|\Delta d_n|$ perturbation on 3D modeling. (A) and (B) Average GDT-HA and lDDT scores of the AS models as a function of their average Cα-Cα PCC_{MODEL} values (see the Methods section). (C) and (D) Similar data obtained for the multiple-templates AM models. Blue triangles represent the scores obtained by applying the template-weighting scheme described in [34] instead of the OL scheme applied for the rest of the data. In (A) through (D), the dashed horizontal lines represent the average quality scores obtained by the default MODELLER.

Effects on multiple-template modeling. Next, we performed perturbation experiments with multiple-template models (see **Fig 5C** and **5D**). Again, the average quality decreases as perturbation increases.

However, when $|\Delta d_n|$ values are fully perturbed, the average GDT-HA now becomes 5.3% lower than the one obtained using the default MODELLER. This behavior is explained by the fact that in perturbation experiments the OL template weighting scheme was employed. When this scheme is applied with optimal (or near-optimal) σ values, it boosts 3D modeling quality, but when it is applied with σ values being weakly correlated with $|\Delta d_n|$ values, it has a detrimental effect (since for each HDDR it uses only the contribution of a randomly chosen template, while the contribution from the best template is likely to be suppressed). In order to make modeling quality less sensitive to errors in $|\Delta d_n|$ estimations, the template weighting scheme of RosettaCM [34] was adopted. In this scheme, the template with the lowest σ value is still assigned the highest weight, but also other templates are given non-zero weights. Using this scheme with a parameter k = 50.0 makes modeling quality more tolerant with respect to the amount of $|\Delta d_n|$ perturbation (see **Fig 5C**).

This data shows that if we were able to predict $|\Delta d_n|$ values with sufficiently high accuracy, the performance of MODELLER would greatly increase. In single-template modeling, obtaining predictions with a PCC of ~0.6 would lead to an increase in GDT-HA of ~2.0%, while in multiple-template modeling, the same PCC would increase GDT-HA by ~11.0% (when using a template-weighting scheme possessing robustness with respect to errors in $|\Delta d_n|$ estimations, such as the RosettaCM scheme).

Modifying the objective function of MODELLER with statistical potential terms

Effect on single-template modeling. In order to identify the optimal way to incorporate the DOPE potential within MODELLER, we performed benchmarks with the AS single-template models by tuning w_{SP} values from 0.1 to 3.5 and by employing HDDRs bearing either σ_{MOD} or $|\Delta d_n|$ values. **Fig 6A** to **6C** show that, with both types of σ, the inclusion of DOPE leads to improvements in 3D modeling. Strikingly, depending on the type of σ, the amount of improvement and the best w_{SP} vary greatly.

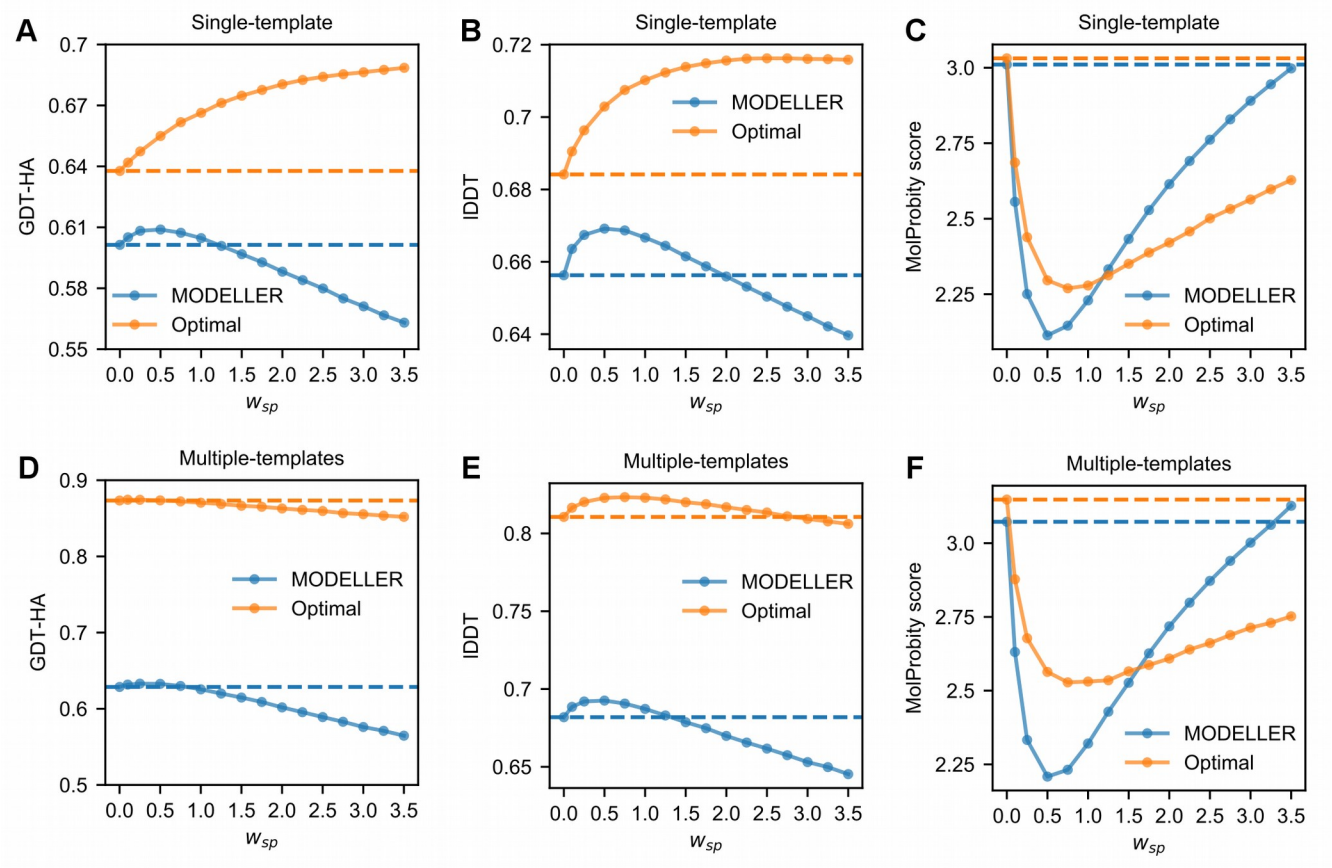


Fig 6. Average quality scores of the models of the analysis set as a function of the w_{SP} with which the DOPE potential has been included in the objective function of MODELLER. (A) to (C) Quality scores of the AS models. (D) to (F) Quality scores of the AM models. (A) through (F) The horizontal dashed lines correspond to the scores obtained when modeling with MODELLER-generated (blue color) or optimal (orange) HDDRs without the use of DOPE.

With σ_{MOD} values, the maximum increase in GDT-HA is observed with a w_{SP} of 0.5. As shown in **Table 3**, when employing DOPE with this w_{SP}, the average GDT-HA improves by a statistically significant 1.3% with respect to the default MODELLER. At the same time, the average lDDT score increases by

2.0%, showing that the use of DOPE also aids local modeling. Of note, when applying DOPE along with the *slow* MDSA protocol, an additional improvement is obtained: the average GDT-HA and lDDT scores now increase by 1.6% and 2.8%.

Table 3. 3D modeling qualities of the AS single-template models built by including DOPE in the objective function of MODELLER.

Strategy	GDT-HA	lDDT	MolProbity score
MODELLER	0.6014 (-)	0.6563 (-)	3.0104 (-)
OPTIMAL	0.6377 (+6.0%)*	0.6842 (+4.2%)*	3.0311 (+0.7%)*
MODELLER-DOPE-0.5[a]	0.6089 (+1.3%)*	0.6692 (+2.0%)*	2.1138 (-29.8%)*
MODELLER-SLOW-DOPE-0.5	0.6112 (+1.6%)*	0.6746 (+2.8%)*	2.0344 (-32.4%)*
MODELLER-DOPE-3.5	0.5631 (-6.4%)*	0.6397 (-2.5%)*	2.9977 (-0.4%)
OPTIMAL-DOPE-0.5	0.6549 (+8.9%)*	0.7029 (+7.1%)*	2.2960 (-23.7%)*
OPTIMAL-DOPE-3.5	0.6885 (+14.5%)*	0.7158 (+9.1%)*	2.6280 (-12.7%)

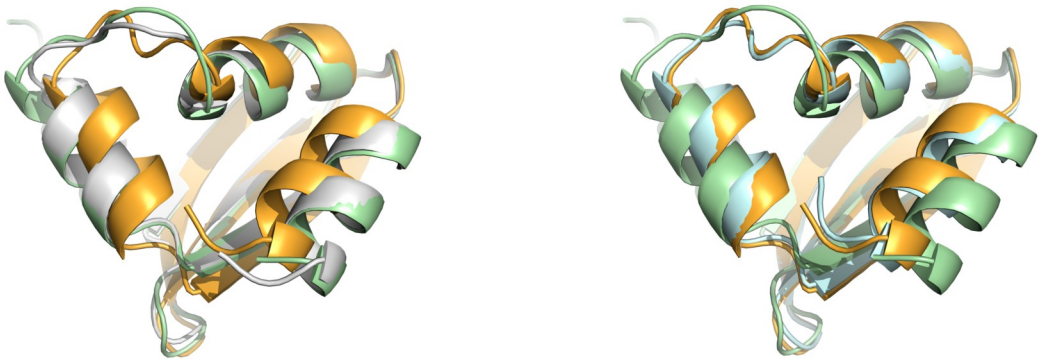
See **Table 1** for the description of contents, columns and most modeling strategies names. See **S3 Table** for a full list of the numerical p-values. [a]The "DOPE-X.X" suffix indicates the use of DOPE with a w_{SP} of X.X.

When modeling with $|\Delta d_n|$ values, the best results are instead obtained with a w_{SP} of 3.5. In this case, DOPE increases the average GDT-HA and lDDT scores by 8.0% and 4.6% with respect to the scores obtained with the same restraints and the standard objective function of MODELLER. The increments in these two metrics are extremely large if computed with respect to the default MODELLER protocol (14.5% and 9.1%). **Fig 7** shows that with the default MODELLER, secondary structure elements that

457 show divergence in the target and template structures are most often modeled in the template
458 conformation. By using optimal HDDRs and DOPE, it is common to see these elements shifting
459 towards target conformations.

460

461



462 **Fig 7. Effects on 3D modeling of optimal σ values and DOPE.** Effects brought by the use $|\Delta d_n|$
463 values and DOPE (with a w_{SP} of 3.5) on the 3D modeling of target *1yd0_chain_A* (colored in orange)
464 using as a template *1yd6_chain_D* (pale green). In the model built using the default MODELLER
465 (colored in white, superposed to its target and template on the left image) the three helices shown in the
466 image are positioned in the same conformation of the template. In the model built employing $|\Delta d_n|$
467 values and DOPE with a w_{SP} of 3.5 (pale cyan, shown on the right) the helices are repositioned in a
468 native-like conformation. Figures rendered with PyMOL [37].

469

Remarkably, the same w_{SP} of 3.5 leads to a large decrease in modeling quality when DOPE is applied along with σ_{MOD} values: in this case, the average GDT-HA and lDDT scores decrease by a large 6.4% and 2.5% with respect to the score obtained without using DOPE.

This data shows that in single-template modeling, the addition of DOPE is much more effective with $|\Delta d_n|$ values than with σ_{MOD} values. Additional insights into this behaviour were provided by the analysis of DOPE energetic landscapes. **Fig 8** shows the representative case of the *1lam_chain_A* and *1dk8_chain_A* targets, where the DOPE energies of models are plotted as a function of their GDT-HA scores. When using single-template HDDRs with σ_{MOD} values, applying DOPE with increasingly high w_{SP} values leads to a decrease in both GDT-HA and DOPE energies. These energies eventually become even lower than the native target structure one. It seems that in the DOPE landscape, near-native conformations are not at an absolute minimum. On the other hand, when modeling with single-template optimal HDDRs, increasing w_{SP} values leads to improvements in GDT-HA while maintaining DOPE energies relatively high. Similar trends are observed in the landscapes of almost all AS models. We speculate that this behaviour is caused by the fact that optimal HDDRs strongly restrain those regions of models which are structurally conserved between the native structures and templates, while they weakly restrain divergent regions. This probably allows to pinpoint the effect of DOPE in the divergent regions (where its addition likely improves modeling over the use of the standard MODELLER objective function) and to keep “rigid” the conserved regions (which are already extremely well-modeled and where DOPE can hardly improve the situation), thus giving rise to a synergistic effect.

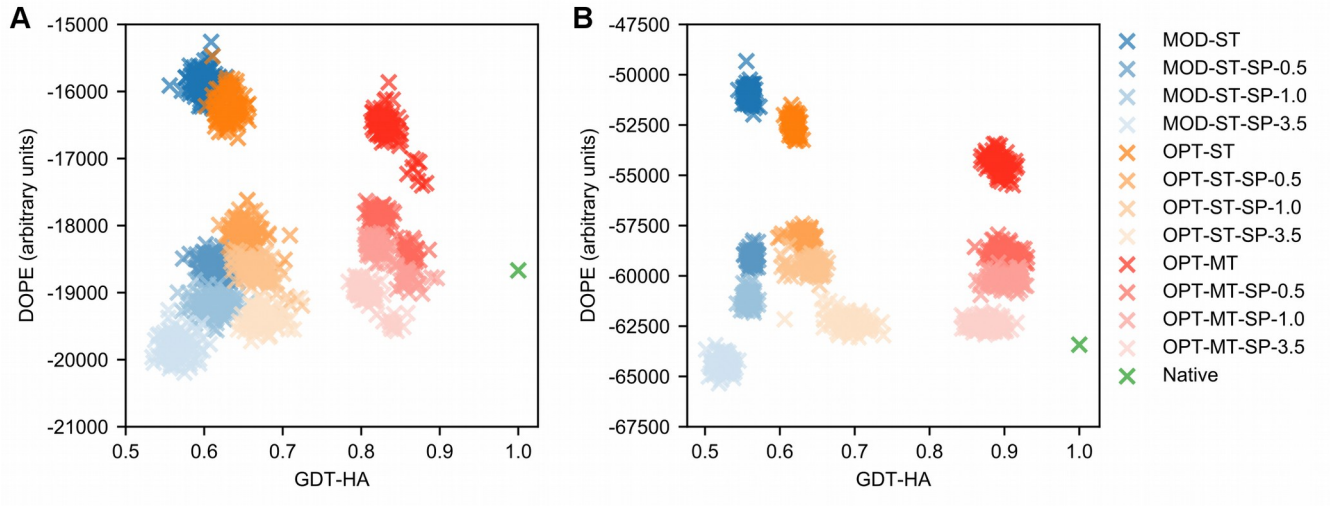


Fig 8. DOPE energy landscapes. DOPE energy landscapes for target (A) *1dk8_chain_A* and (B) *1lam_chain_A* modeled using different strategies. 100 decoys were built for each strategy and their GDT-HA scores are plotted here against their DOPE energies. The strategies with the "MOD-ST" prefix adopted MODELLER-generated HDDRs and a single template (blue-shaded dots), those with the "OPT-ST" prefix adopted optimal HDDRs and a single template (orange-shaded dots) and those with the "OPT-MT" prefix adopted optimal HDDRs and multiple templates (red-shaded dots). The "SP-X.X" suffix indicates the use of DOPE with a w_{SP} of X.X. The green dots correspond to the DOPE-minimized native target structure.

Effect on multiple-template modeling. Next, we explored the effect of DOPE in multiple-template modeling (see **Fig 6D** to **6F**). The trend observed when employing MODELLER-generated restraints is reminiscent of the single-template modeling one, although the improvements are slightly smaller. **Table 4** shows that the best w_{SP} is 0.5, which results in an average increase in GDT-HA and lDDT of 0.6% and 1.6% with respect to the scores obtained with the default MODELLER. By employing DOPE with this w_{SP} along with the *slow* MDSA protocol, an additional improvement can be reached: the average GDT-

HA and lDDT scores now improve by 1.0% and 2.2%. When further increasing w_{SP}, we assist to a decrease in 3D modeling qualities.

Table 4. 3D modeling qualities of the AM multiple-template models built by including DOPE in the objective function of MODELLER.

Strategy	GDT-HA	lDDT	MolProbity score
MODELLER	0.6287 (-)	0.6819 (-)	3.0725 (-)
OPTIMAL	0.8733 (+38.9%)*	0.8106 (+18.9%)*	3.1478 (+2.4%)
MODELLER-DOPE-0.5	0.6327 (+0.6%)*	0.6926 (+1.6%)*	2.2086 (-28.1%)*
MODELLER-SLOW-DOPE-0.5	0.6347 (+1.0%)*	0.6971 (+2.2%)*	2.1152 (-31.2%)*
MODELLER-DOPE-3.5	0.5646 (-10.2%)*	0.6453 (-5.4%)*	3.1267 (+1.8%)*
OPTIMAL-DOPE-0.5	0.8736 (+39.0%)*	0.8229 (+20.7%)*	2.5635 (-16.6%)*
OPTIMAL-DOPE-3.5	0.8519 (+35.5%)*	0.8061 (+18.2%)*	2.7520 (-10.4%)*

See **Table 1** for the description of contents, columns and most modeling strategies names. See **S4 Table** for a full list of the numerical p-values.

The results observed when combining DOPE with optimal multiple-template HDDRs are different. No value of w_{SP} is able to bring a relevant improvement in GDT-HA. As w_{SP} increases over 1.0, the scores even start to decrease in a significant way, although it seems that DOPE is able to bring at least a small improvement in lDDT.

This counterintuitive behaviour can in part be explained from the analysis of DOPE energy landscapes. **Fig 8** shows that when using optimal multiple-template HDDRs, the quality of models is already higher

than the one obtained with optimal single-template HDDRs. In this case, applying large w_{SP} values leads to a decrease in DOPE energies and GDT-HA. The plots show that the models built with optimal HDDRs seem to be attracted towards a local energy minimum of DOPE, which does not correspond to the native state, but is located relatively near it. Therefore, when using optimal restraints, minimizing the DOPE of a structure distant from the native state (like in the case of single-template modeling), tends to increase its GDT-HA, but when the structure is already very close to the native state (such as in the case of multiple-template modeling), it tends to decrease its GDT-HA.

Effects on stereochemical quality. In terms of stereochemichal quality, the use of DOPE seems to be highly beneficial in both single and multiple-template modeling and with both MODELLER-generated and optimal HDDRs (see **Fig 6, Table 3** and **4**). For example, when employing σ_{MOD} values and DOPE with a w_{SP} of 0.5, the average MolProbity score of the AS models decreases by a large 29.8% with respect to the default MODELLER. Additional improvements in MolProbity scores are observed when coupling DOPE to the *slow* MDSA protocol. We found that the MolProbity score component in which DOPE brings the largest improvement is by far the "Clash Score", meaning that the potential helps to remove steric clashes from models. Therefore, the inclusion of DOPE in the objective function of MODELLER represents a fast and effective way of improving the stereochemical quality of its models. This approach increases computational times by a factor of ~6.5 when employing the *very_fast* MDSA protocol (and ~16.5 with the *slow* protocol), but on modern hardware the default MODELLER algorithm usually takes a few seconds to complete a model, therefore in absolute terms the model building process is still relatively fast.

Comparison between DOPE and DFIRE in 3D modeling. We also tested the effect of adding DFIRE in the objective function of MODELLER. Overall, DFIRE seems to have very similar effects to the ones described for DOPE (see **S3 Table**, **S4 Table** and **S5 Fig**), because their terms have very similar

forms (see **S3 FigB**). However, when modeling with σ_{MOD} values, DOPE seems to slightly outperform DFIRE in terms of all-atom local quality (expressed by lDDT scores). When using a w_{SP} of 0.5 and σ_{MOD} values, DOPE yields for the AS models an average lDDT score 0.5% higher than the one obtained with DFIRE, a small but statistically significant improvement (Wilcoxon signed-rank test, p-value = 4.6e-35). Therefore, we suggest that in MODELLER, DOPE should be preferred over DFIRE.

Discussion

Improving the quality of HM predictions is clearly an area of great relevance in Biomedical Research [38], given that the applicability of this methodology is expected to increase in the next years [29]. Right now, a large portion of targets can be modeled only with low accuracy, due to the remote homology relationship (under 30% SeqId) with their templates. A solution to this problem could potentially come from advances in 3D model building or refinement algorithms. In this work, we have explored two main promising strategies to increase the accuracy of the original MODELLER algorithm.

The use of optimal σ values (that is, $|\Delta d_n|$ values) greatly increases the 3D modeling quality of the program. Since $|\Delta d_n|$ values can only be obtained by knowing the exact amount of divergence between the structure of a target and its templates, they can not be used in real-life protein structure prediction scenarios (where the target structure is of course unknown).

However, as first shown by the Lee group [19], $|\Delta d_n|$ values may be estimated through a machine learning system. These authors developed a random forest which obtained estimations with an average Cα-Cα PCC of ~0.35. The use of this predictor led to only a very small improvement in terms of 3D modeling quality. Our data (which describes the relationship between 3D modeling quality and errors in $|\Delta d_n|$ estimations) shows that increasing the PCC of a similar predictor by at least 0.2-0.3 units could translate in a significant improvement of MODELLER.

The other strategy that we have investigated is the inclusion of statistical potential terms, such as DOPE, in the objective function of MODELLER. We show that employing such potentials in the 3D model building phase of MODELLER robustly increases 3D modeling quality and provides a fast and effective way to improve the stereochemical details models. In order to allow the user community of MODELLER to deploy this strategy in their modeling pipelines, we share the Python code implementing it. In future research, it will be interesting to see if there exist potentials with an even more beneficial effect on 3D model building in MODELLER.

Our results have implications also for other Structural Bioinformatics tools. RosettaCM and I-TASSER borrow from MODELLER the use of HDDRs [34, 39-40] and programs like MULTICOM [41] and Pcons [42] implement MODELLER at some point in their protein modeling pipelines. The strategies presented in this work can certainly be implemented in these protocols to improve their quality.

Of note, in the protein structure refinement field, restraints are built from a starting model and the aim is to guide the model towards its native conformation [43]. While in the HM context we may estimate $|\Delta d_n|$ values between a target native structure and a template, in protein structure refinement they could be similarly estimated between a native structure and its unrefined model. Methods to predict the local accuracy of 3D models already reach good performances [44]. It is reasonable to think that with a sufficiently accurate predictor, the $|\Delta d_n|$ prediction strategy could also lead to improvements in current refinement strategies.

The development of deep learning techniques [45] has recently brought advances in the field of contact and distance map prediction [46]. We suggest that such methodologies could be well adapted to the problem of $|\Delta d_n|$ estimation. In future studies, we will concentrate on using this type approach to tackle the problem of σ values assignment. Since a machine learning model usually performs predictions in a relatively small amount of time, the $|\Delta d_n|$ estimation approach has the potential to greatly improve the

588 "modeling by satisfaction of spatial restraints" strategy of MODELLER at the price of small
589 computational cost.

590 Acknowledgements

591 The authors wish to acknowledge Fabio Mastrantuono and Fransceso Pesce for helpful discussions and
592 their precious help.

593 This work is dedicated to the memory of our beloved mentor Prof. Francesco Bossa.

594 References

595 1. Rigden DJ, editor. From Protein Structure to Function with Bioinformatics. 2nd ed. Springer
596 Netherlands; 2017.

597 2. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of
598 methods of protein structure prediction (CASP)-Round XII. Proteins. 2018;86 Suppl 1: 7–15.
599 doi:10.1002/prot.25415

600 3. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and
601 PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res.
602 1997;25: 3389–3402.

603 4. Söding J. Protein homology detection by HMM-HMM comparison. Bioinformatics. 2005;21:
604 951–960. doi:10.1093/bioinformatics/bti125

605 5. Yan R, Xu D, Yang J, Walker S, Zhang Y. A comparative assessment and analysis of 20
606 representative sequence alignment methods for protein structure prediction. Sci Rep. 2013;3:
607 2619. doi:10.1038/srep02619

6. Kryshtafovych A, Monastyrskyy B, Fidelis K, Moult J, Schwede T, Tramontano A. Evaluation of the template-based modeling in CASP12. Proteins. 2018;86 Suppl 1: 321–334. doi:10.1002/prot.25425

7. Meier A, Söding J. Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling. PLoS Comput Biol. 2015;11: e1004343. doi:10.1371/journal.pcbi.1004343

8. Park H, Ovchinnikov S, Kim DE, DiMaio F, Baker D. Protein homology model refinement by large-scale energy optimization. Proc Natl Acad Sci USA. 2018;115: 3054–3059. doi:10.1073/pnas.1719115115

9. Heo L, Feig M. Experimental accuracy in protein structure refinement via molecular dynamics simulations. Proc Natl Acad Sci USA. 2018;115: 13276–13281. doi:10.1073/pnas.1811364115

10. Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinformatics. 2016;54: 5.6.1-5.6.37. doi:10.1002/cpbi.3

11. Wallner B, Elofsson A. All are not equal: a benchmark of different homology modeling programs. Protein Sci. 2005;14: 1315–1327. doi:10.1110/ps.041253405

12. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993;234: 779–815. doi:10.1006/jmbi.1993.1626

13. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, et al. CHARMM: the biomolecular simulation program. J Comput Chem. 2009;30: 1545–1614. doi:10.1002/jcc.21287

14. Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol. 2017; doi:10.1016/j.jmb.2017.12.007

15. Joo K, Lee J, Sim S, Lee SY, Lee K, Heo S, et al. Protein structure modeling for CASP10 by multiple layers of global optimization. Proteins. 2014;82 Suppl 2: 188–195. doi:10.1002/prot.24397

16. Joo K, Joung I, Lee SY, Kim JY, Cheng Q, Manavalan B, et al. Template based protein structure modeling by global optimization in CASP11. Proteins. 2016;84 Suppl 1: 221–232. doi:10.1002/prot.24917

17. Hong SH, Joung I, Flores-Canales JC, Manavalan B, Cheng Q, Heo S, et al. Protein structure modeling and refinement by global optimization in CASP12. Proteins. 2018;86 Suppl 1: 122–135. doi:10.1002/prot.25426

18. Joo K, Lee J, Seo J-H, Lee K, Kim B-G, Lee J. All-atom chain-building by optimizing MODELLER energy function using conformational space annealing. Proteins. 2009;75: 1010–1023. doi:10.1002/prot.22312

19. Lee J, Lee K, Joung I, Joo K, Brooks BR, Lee J. Sigma-RF: prediction of the variability of spatial restraints in template-based modeling by random forest. BMC Bioinformatics. 2015;16: 94. doi:10.1186/s12859-015-0526-z

20. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci. 2002;11: 2714–2726. doi:10.1110/ps.0217002

649 21. Lee J, Lee J, Sasaki TN, Sasai M, Seok C, Lee J. De novo protein structure prediction by
650 dynamic fragment assembly and conformational space annealing. Proteins. 2011;79: 2403–
651 2417. doi:10.1002/prot.23059

652 22. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential
653 improves prediction of specificity and structure for proteins and protein-protein complexes. J
654 Mol Biol. 2003;326: 1239–1259.

655 23. Shen M-Y, Sali A. Statistical potential for assessment and prediction of protein structures.
656 Protein Sci. 2006;15: 2507–2524. doi:10.1110/ps.062416606

657 24. Larsson P, Wallner B, Lindahl E, Elofsson A. Using multiple templates to improve quality of
658 homology models in automated homology modeling. Protein Sci. 2008;17: 990–1002.
659 doi:10.1110/ps.073344908

660 25. Wang G, Dunbrack RL. PISCES: recent improvements to a PDB sequence culling server.
661 Nucleic Acids Res. 2005;33: W94-98. doi:10.1093/nar/gki402

662 26. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score.
663 Nucleic Acids Res. 2005;33: 2302–2309. doi:10.1093/nar/gki524

664 27. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template
665 quality. Proteins. 2004;57: 702–710. doi:10.1002/prot.20264

666 28. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5?
667 Bioinformatics. 2010;26: 889–895. doi:10.1093/bioinformatics/btq066

668 29. Schwede T. Protein modeling: what happened to the “protein structure gap”? Structure.
669 2013;21: 1531–1540. doi:10.1016/j.str.2013.08.007

30. Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, et al. CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Res. 2017;45: D289–D295. doi:10.1093/nar/gkw1098

31. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2011;9: 173–175. doi:10.1038/nmeth.1818

32. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics. 2013;29: 2722–2728. doi:10.1093/bioinformatics/btt473

33. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr. 2010;66: 12–21. doi:10.1107/S0907444909042073

34. Thompson J, Baker D. Incorporation of evolutionary information into Rosetta comparative modeling. Proteins. 2011;79: 2380–2388. doi:10.1002/prot.23046

35. Rykunov D, Fiser A. New statistical potential for quality assessment of protein models and a survey of energy functions. BMC Bioinformatics. 2010;11: 128. doi:10.1186/1471-2105-11-128

36. Chopra G, Kalisman N, Levitt M. Consistent refinement of submitted models at CASP using a knowledge-based potential. Proteins. 2010;78: 2668–2678. doi:10.1002/prot.22781

37. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. 2015.

38. Schwede T, Sali A, Honig B, Levitt M, Berman HM, Jones D, et al. Outcome of a workshop on applications of protein models in biomedical research. Structure. 2009;17: 151–159. doi:10.1016/j.str.2008.12.014

691 39. Song Y, DiMaio F, Wang RY-R, Kim D, Miles C, Brunette T, et al. High-resolution comparative
692 modeling with RosettaCM. Structure. 2013;21: 1735–1742. doi:10.1016/j.str.2013.08.005

693 40. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and
694 function prediction. Nat Methods. 2015;12: 7–8. doi:10.1038/nmeth.3213

695 41. Wang Z, Eickholt J, Cheng J. MULTICOM: a multi-level combination approach to protein
696 structure prediction and its assessments in CASP8. Bioinformatics. 2010;26: 882–888.
697 doi:10.1093/bioinformatics/btq058

698 42. Wallner B, Fang H, Elofsson A. Automatic consensus-based fold recognition using Pcons,
699 ProQ, and Pmodeller. Proteins. 2003;53 Suppl 6: 534–541. doi:10.1002/prot.10536

700 43. Feig M. Computational protein structure refinement: Almost there, yet still so far to go. Wiley
701 Interdiscip Rev Comput Mol Sci. 2017;7. doi:10.1002/wcms.1307

702 44. Uziela K, Menéndez Hurtado D, Shu N, Wallner B, Elofsson A. ProQ3D: improved model
703 quality assessments using deep learning. Bioinformatics. 2017;33: 1578–1580.
704 doi:10.1093/bioinformatics/btw819

705 45. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521: 436–444.
706 doi:10.1038/nature14539

707 46. Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin AMJJ. Assessment of contact
708 predictions in CASP12: Co-evolution and deep learning coming of age. Proteins. 2018;86 Suppl
709 1: 51–66. doi:10.1002/prot.25407

Supporting Information

S1 Table. Physical terms of the MODELLER objective function. Note how by default the objective function does not include any "physical" attractive term between non-bonded atoms (Lennard-Jones and Coulomb potential terms from CHARMM22 [Brooks at al., 2009] are missing). The only attractive terms in the objective function are homology-derived distance restraints (see **S2 Table**).

S2 Table. Homology-derived terms of the MODELLER objective function.

S3 Table. 3D modeling qualities of the AS single-template models built with different modeling strategies. See **Table 1** in the main text for the description of contents, columns and most modeling strategies names.

S4 Table. 3D modeling qualities of the AM multiple-templates models built with different modeling strategies. See **Table 1** and **2** in the main text and **S3 Table** for the description of contents, columns and most modeling strategies names.

S1 Fig. Properties of the analysis set. (A) SeqId histogram of the pairwise target-template alignments in the AS models obtained using TM-align and HHalign. (B) Target coverage histograms of the same alignments. (C) Chain length histograms of the 225 AS targets, the 118 AM targets and all the 472 template chains of the analysis set. (D) CATH classes frequencies of the AS and AM targets compared to those in the entire CATH 4.2.0 database [Dawson et al., 2017].

S2 Fig. Average PCC_{MODEL} values in $|\Delta d_n|$ perturbation experiments plotted as a function of f_e. Data for the four HDDRs groups of MODELLER is shown. As f_e (that is, the fraction of perturbed $|\Delta d_n|$ values in models) increases, the average correlation between $|\Delta d_n|$ values and their perturbed counterpart decreases. (A) AS models. (B) AM models.

S3 Fig. Analysis of the terms of the DOPE and DFIRE potentials. (A) Forms of the 12561 terms of DOPE [Shen and Sali, 2006]. Each term is associated to a couple of heavy atom types from the 20 standard residues. Irrespective of the atom types, all the functions start to acquire a flat shape above the 8.0 Å threshold. (B) Confrontation of DOPE and DFIRE [Zhou and Zhou, 2002] terms. An hexbin density plot compares 364269 data points from all the 12561 terms of DOPE (x-axis) and DFIRE (y-axis) (each term has 29 points, which report the score of the potential in a linear space from 0.75 to 14.75 Å). The scores of the two potentials are highly correlated (Pearson correlation coefficient = 0.99).

S4 Fig. Accuracy of the pairwise target-template HHalign alignments of the AS models. The x-axis reports the SeqId between the target and template sequences in TM-align alignments. The y-axis reports the accuracy of the corresponding HHalign alignment. The accuracy is computed as the ratio H_m/T_m, where T_m is the total number of matches in the TM-align alignment and H_m is the number of "correct" matches in HHalign alingments (that is, those HHalign matches which are also found in the TM-align alignment). The average accuracy is 0.87.

S5 Fig. Average quality scores of the analysis set models as a function of the w_{sp} value with which the DFIRE or DOPE statistical potentials have been included in the objective function of MODELLER. The horizontal dashed lines correspond to the scores obtained when modeling with MODELLER-generated (blue color) or optimal (orange) HDDRs without the use of statistical potentials. (A) to (C) quality scores of the AS models. (D) to (F) quality scores of the AM models.

S1 Text. Description of the GDT-HA and lDDT metrics for model quality evaluation.

S2 Text. Obtaining optimal parameters for single-template HDDRs.

S3 Text. Obtaining optimal parameters for multiple-template HDDRs.