BI-INTERPRETATION IN WEAK SET THEORIES

JOEL DAVID HAMKINS AND ALFREDO ROQUE FREIRE

Abstract. In contrast to the robust mutual interpretability phenomenon in set theory, Ali Enayat proved that bi-interpretation is absent: distinct theories extending ZF are never bi-interpretable and models of ZF are bi-interpretable only when they are isomorphic. Nevertheless, for natural weaker set theories, we prove, including Zermelo-Fraenkel set theory ZFC− without power set and Zermelo set theory Z, there are nontrivial instances of bi-interpretation. Specifically, there are well-founded models of ZFC− that are bi-interpretable, but not isomorphic—even ⟨H_{ω1}, ∈⟩ and ⟨H_{ω2}, ∈⟩ can be bi-interpretable—and there are distinct bi-interpretable theories extending ZFC−. Similarly, using a construction of Mathias, we prove that every model of ZF is bi-interpretable with a model of Zermelo set theory in which the replacement axiom fails.

1. Introduction

Set theory exhibits a robust mutual interpretability phenomenon: in a given model of set theory, we can define diverse other interpreted models of set theory. In any model of Zermelo-Fraenkel ZF set theory, for example, we can define an interpreted model of ZFC+GCH, via the constructible universe, as well as definable interpreted models of ZF + ¬AC, of ZFC + MA + ¬CH, of ZFC + b < ℵ, and so on for hundreds of other theories. For these latter theories, set theorists often use forcing to construct outer models of the given model; but nevertheless the Boolean ultrapower method provides definable interpreted models of these theories inside the original model (see theorem 7). Similarly, in models of ZFC with large cardinals, one can define fine-structural canonical inner models with large cardinals and models of ZF satisfying various determinacy principles, and vice versa. In this way, set theory exhibits an abundance of natural mutually interpretable theories.

Do these instances of mutual interpretation fulfill the more vigorous conception of bi-interpretation? Two models or theories are mutually interpretable, when merely each is interpreted in the other, whereas bi-interpretation requires that the interpretations are invertible in a sense after iteration, so that if one should interpret one model or theory in the other and then re-interpret the first theory inside that, then the resulting model should be definably isomorphic to the original universe (precise definitions in sections §2 and §3). The interpretations mentioned above are not bi-interpretations, for if we start in a model of ZFC + ¬CH and then go to L in order to interpret a model of ZFC + GCH, then we’ve already discarded too much set-theoretic information to expect that we could get a copy of our original model back by interpreting inside L. This problem is inherent, in light of the following

This research was supported by grant 2017/21020-0, São Paulo Research Foundation (FAPESP). The research project grew out of the second author’s PhD dissertation [Fre19], with related philosophical work in [Fre08, FF19]. Commentary can be made about this article on the first author’s blog at http://jdh.hamkins.org/bi-interpretation-in-weak-set-theories.
theorem of Ali Enayat, showing that indeed there is no nontrivial bi-interpretation phenomenon to be found amongst the set-theoretic models and theories satisfying ZF. In interpretation, one must inevitably discard set-theoretic information.

**Theorem 1** (Enayat [Ena16]).

1. **ZF** is solid: no two models of **ZF** are bi-interpretable.
2. **ZF** is tight: no two distinct theories extending **ZF** are bi-interpretable.

The proofs of these theorems, provided in section §6, seem to use the full strength of ZF, and Enayat had consequently inquired whether the solidity/tightness phenomenon somehow required the strength of ZF set theory. In this paper, we shall find support for that conjecture by establishing nontrivial instances of bi-interpretation in various natural weak set theories, including Zermelo-Fraenkel theory **ZFC**\(^-\), without the power set axiom, and Zermelo set theory **Z**, without the replacement axiom.

**Main Theorems.**

1. **ZFC**\(^-\) is not solid: there are well-founded models of **ZFC**\(^-\) that are bi-interpretable, but not isomorphic.
2. Indeed, it is relatively consistent with **ZFC** that \(H_{\omega_1}, \in\) and \(H_{\omega_2}, \in\) are bi-interpretable.
3. **ZFC**\(^-\) is not tight: there are distinct bi-interpretable extensions of **ZFC**\(^-\).
4. **Z** is not solid: there are well-founded models of **Z** that are bi-interpretable, but not isomorphic.
5. Indeed, every model of **ZF** is bi-interpretable with a transitive inner model of **Z** in which the replacement axiom fails.
6. **Z** is not tight: there are distinct bi-interpretable extensions of **Z**.

These claims are made and proved in theorems 20, 17, 21, and 22. We shall in addition prove the following theorems on this theme:

7. Well-founded models of **ZF** set theory are never mutually interpretable.
8. The Väänänen internal categoricity theorem does not hold for **ZFC**\(^-\), not even for well-founded models.

These are theorems 14 and 16. Statement (8) concerns the existence of a model \(\langle M, \in, \bar{\in}\rangle\) satisfying **ZFC**\(^-\)\((\in, \bar{\in})\), meaning **ZFC**\(^-\) in the common language with both predicates, using either \(\in\) or \(\bar{\in}\) as the membership relation, such that \(\langle M, \in\rangle\) and \(\langle M, \bar{\in}\rangle\) are not isomorphic.

### 2. Interpretability in models

Let us briefly review what interpretability means. The reader is likely familiar with the usual interpretation of the complex field \(\langle \mathbb{C}, +, \cdot, 0, 1, i\rangle\) in the real field \(\langle \mathbb{R}, +, \cdot\rangle\), where one represents the complex number \(a + bi\) with the pair of real numbers \((a, b)\). The point is that complex number field operations are definable operations on these pairs in the real field. Conversely, it turns out that the real field is not actually interpreted in the complex field, but it is interpreted in the complex field equipped also with the complex conjugate operation \(z \mapsto \bar{z}\), for in this case you can define the real line as the class of \(z\) for which \(z = \bar{z}\). The reader is also likely familiar with the usual interpretation of the rational field \(\langle \mathbb{Q}, +, \cdot\rangle\) in the integer ring \(\langle \mathbb{Z}, +, \cdot\rangle\), where one represents rational numbers as equivalence classes of integer pairs \((p, q)\), written more conveniently as fractions \(\frac{p}{q}\), with \(q \neq 0\), considered
under the equivalence relation \( \frac{p}{q} \equiv \frac{r}{s} \leftrightarrow ps = qr \); one then defines the rational field structure by means of the familiar fractional arithmetic. Another example would be the structure of hereditarily finite sets \( \langle \text{HF}, \in \rangle \), which is interpreted in the standard model of arithmetic \( \langle \mathbb{N}, +, \cdot, 0, 1, < \rangle \) by the Ackermann relation, the relation for which \( n E m \) just in case the \( n \)th binary digit of \( m \) is 1; this relation is definable in arithmetic and the reader may verify that \( \langle \text{HF}, \in \rangle \cong \langle \mathbb{N}, E \rangle \).

More generally, we interpret one model in another as follows.

**Definition 2.** A model \( N = \langle N, R, \ldots \rangle \) in a first-order language \( L_N \) is interpreted in another model \( M \) in language \( L_M \), if for some finite \( k \) there is an \( M \)-definable class \( N^* \subseteq M^k \) of \( k \)-tuples and \( M \)-definable relations \( R^{N^*} \) on \( N^* \) for relation symbols \( R \) in \( L_N \), as well as functions (defined via their graphs) and constants, and an \( M \)-definable equivalence relation \( \simeq \) on \( N^* \), which is a congruence with respect to those relations, such that \( N \) is isomorphic to the quotient structure \( \langle N^*, R^{N^*}, \ldots \rangle/\simeq \).

![Figure 1. Model N is interpreted in model M](image1)

In models supporting an internal encoding of finite tuples, such as the models of arithmetic or set theory, there is no need for \( k \)-tuples and one may regard \( N \subseteq M \) directly. In models of arithmetic or models of set theory with a definable global well-order, there is no need for the equivalence relation \( \simeq \), since one can pick canonical least representatives from each class. Theorem 8 shows furthermore that in models of ZF, even when there isn’t a definable global well-order, one can still omit the need for the equivalence relation \( \simeq \) by means of Scott’s trick: replace every equivalence class with the set of its \( \in \)-minimal rank elements. This trick doesn’t necessarily work, however, in models of ZFC\(^-\), without the power set axiom, since in this theory the class of minimal rank members could still be a proper class, although in ZF it is always a set.

**Definition 3.** Models \( M \) and \( N \) are **mutually interpretable**, if each of them is interpreted in the other.

![Figure 2. Models M and N are mutually interpreted](image2)
Since (the quotient of) $N^*$ is isomorphic to $N$, we may find the corresponding copy of $M$ inside it, and similarly we may find the copy of $N$ inside $M^*$, as illustrated below, where we have now suppressed the representation of the equivalence relation.

\textbf{Figure 3.} Iterating the interpretations of models $M$ and $N$

One may simply compose the isomorphisms $i$ and $j$ to form an isomorphism $ji$ between $M$ and $M^*$, which is now a definable subset of $M$ (although the map $ji$ may not be definable). Because in the general case these are quotient structures, one should think of the map $ji$ as a relation rather than literally a function, since each object $a$ in $M$ is in effect associated with an entire equivalence class $\bar{a}$ in the quotient structure of $M$. Similarly, the composition $ij$ is an isomorphism of the structure $N$ with $N^*$, which is now definable in $N$.

It will be instructive to notice that we have little reason in general to expect the map $ji$ to be definable in $M$, and this issue is the key difference between the mutual interpretation of models and their bi-interpretation.

\textbf{Definition 4.} Models $M$ and $N$ are \textit{bi-interpretable}, if they are mutually interpretable in such a way that the isomorphisms $ji : M \cong M^*$ and $ij : N \cong N^*$ arising by composing the interpretation maps are definable in the original models $M$ and $N$, respectively.

\textbf{Figure 4.} Models $M$ and $N$ are bi-interpreted

A somewhat cleaner picture emerges, as indicated below, if we should simply identify the model $N$ with its isomorphic copy $N^*$ and $M^*$ with $M$. Let us now use $i$ to denote the isomorphism of $M$ with its copy $M$ inside $N$, and $j$ is the isomorphism of $N$ with its copy $N$ inside $M$. This picture applies with either mutual interpretation or bi-interpretation, the difference being that in the case of bi-interpretation, the isomorphisms $i$ and $j$ are definable in $M$ and $N$, respectively. Furthermore, by iterating these interpretations, in the case either of mutual
interpretation or bi-interpretation, one achieves a fractal-like nested sequence of definable structures, with isomorphisms at each level. In the case of bi-interpretation, all these maps are definable in $M$, and any of the later maps is definable inside any of the successive copies of $M$ or $N$ in which it resides.

![Figure 5. Iterated bi-interpretations](image)

An even stronger connection between models is the notion of bi-interpretation synonymy. This relation was introduced by Bouvere in [dB65] and recently deepened by Friedman and Visser in [FV14].

**Definition 5.** Models $M$ and $N$ are bi-interpretation synonymous, if there is a bi-interpretation for which (i) the domains of the interpreted structures are in each case the whole structure; and (ii) the equivalence relations used in the interpretation are the identity relations on $M$ and $N$.

In remarks after theorem 15, we explain how every instance of bi-interpretation between models of ZF can be transformed to an instance of bi-interpretation synonymy.

3. Interpretability in theories

Let us now explain how interpretability works in theories, as opposed to models. At bottom, an interpretation of one theory in another amounts to a uniform method of interpreting a model of the first theory in any model of the second theory. Specifically, we interpret one theory $T_1$ in another theory $T_2$, by providing a way to translate the $T_1$ language and structure into the language and structure available in $T_2$, in such a way that the translation of every $T_1$ theorem is provable in $T_2$.

In somewhat more detail, the interpretation $I$ of theory $T_1$ in language $L_1$ into theory $T_2$ in language $L_2$ should provide, first of all, an $L_2$-formula $U(\bar{x})$ that will define the interpreted domain and an $L_2$-expressible relation $\bar{x} = I \bar{y}$, the interpretation of equality, that $T_2$ proves is an equivalence relation on tuples in $U$. Next, for each relation symbol $R$ of $L_1$, the interpretation should provide a translation $R^I$ as an $L_2$ formula of the same arity on $U$ as $R$ has in $L_1$, and $T_2$ should prove that this relation is well-defined modulo $=^I$. Finally, functions are to be handled by considering their graphs $f(\bar{x}) = y$ as definable relations and interpreting these graph relations, but with the proviso that $T_2$ proves that the interpreted relation is
indeed well-defined and functional on $U$ modulo $=^I$. Ultimately, therefore, the theory $T_2$ will prove that $=^I$ is a congruence with respect to the interpreted relations $R^I$ and functions $f^I$.

Having thus interpreted the atomic $L_1$ structure, one may naturally extend the interpretation to all $L_1$ assertions as follows.

$$
\begin{align*}
\varphi \wedge \psi^I &= \varphi^I \wedge \psi^I \\
\neg \varphi^I &= \neg \varphi^L \\
\exists x \varphi(x)^I &= \exists \bar{x} U(\bar{x}) \wedge \varphi(\bar{x})^I
\end{align*}
$$

Finally, for this to be an interpretation of $T_1$ in $T_2$, we insist that $T_2$ prove that $T_1$ holds for the interpreted structure, or in other words, that

$$T_1 \vdash \varphi \quad \text{implies} \quad T_2 \vdash \varphi^I.$$

It is clear that if $T_1$ is interpreted in $T_2$ by interpretation $I$, then in any model $M \models T_2$ we may use the interpretation to define a model $N = \langle U, R^I, f^I, \ldots \rangle / =^I$ of $T_1$. The domain of $N$ will consist of the quotient of $U$ by $=^I$ as $M$ sees it, and the relations $R$ and functions $f$ of $N$, if any, will be exactly those defined by $R^I$ and $f^I$ via the interpretation.

**Definition 6.**

1. Two theories $T_1$ and $T_2$ are *mutually interpretable*, if each of them is interpretable in the other.
2. Two theories $T_1$ and $T_2$ are *bi-interpretable*, in contrast, if they are mutually interpretable by interpretations $I$ and $J$ respectively, which are provably invertible, in that the theory $T_1$ proves that the universe is isomorphic, by a definable isomorphism map, to the model resulting by first interpreting via $J$ to get a model of $T_2$ and then interpreting via $I$ to get a model of $T_1$ inside that model; and similarly the theory $T_2$ proves that its universe is definably isomorphic to the model obtained by first interpreting via $I$ to get a model of $T_1$ and then inside that model via $J$ to get a model of $T_2$.

With mutual interpretation of theories, one can start with a model $M$ of $T_1$ and then use $J$ to define within it a model $N = J^M$ of $T_2$, which can then define a model $M' = I^N$ of $T_1$ again. With mutual interpretability, there is no guarantee that $M'$ and $M$ are isomorphic or even elementarily equivalent, beyond the basic requirement that both are models of $T_1$. The much stronger requirements of bi-interpretability, in contrast, ensure that $M'$ and $M$ are isomorphic, and furthermore isomorphic by a definable isomorphism relation, which the theory $T_1$ proves is an isomorphism. And similarly when interpreting in models of $T_2$. Thus, with bi-interpretation, each theory sees how its own structure is faithfully copied via the definable isomorphisms $f_1$ and $f_2$ under iterations of the interpretations. With bi-interpretability, it therefore follows that

1. For every $L_1$ assertion $\alpha$

$$T_1 \models \forall \bar{x} \ \alpha(x_1, \ldots, x_n) \leftrightarrow \alpha(I(f_1(x_1), f_1(x_2), \ldots, f_1(x_n))).$$

2. For every $L_2$ assertion $\beta$

$$T_2 \models \forall \bar{y} \ \beta(y_1, \ldots, y_m) \leftrightarrow \beta(I(f_2(y_1), f_2(y_2), \ldots, f_2(y_m))).$$
In light of the quotients by the equivalence relations, the functions $f_1$ and $f_2$ are more properly thought of as relations $f_1(x) = \bar{y}$ well-defined with respect to those relations; they need not be functional on points, but only in the quotient.

### 4. Mutual interpretation of diverse set theories

Let us briefly establish the mutual interpretability phenomenon in set theory.

**Theorem 7.** The following theories are pairwise mutually interpretable.

1. ZF
2. ZFC
3. ZFC + GCH
4. ZFC + $V = L$
5. ZF + $\neg$AC
6. ZFC + $\neg$CH
7. ZFC + MA + $\neg$CH
8. ZFC + $b < d$
9. etc. etc. etc.

**Proof.** In many instances, the interpretation is obtained by defining a suitable inner model of the desired theory. For example, in any model of ZF, we can define the constructible universe, and thereby find an interpretation of ZFC + $V = L$ and hence ZFC + GCH and so forth in ZF. Some of the interpretations involve forcing, however, which we usually conceive as a method for constructing outer models, rather than defining models inside the original model. Nevertheless, one can use forcing via the Boolean ultrapower method to define interpreted models of the forced theory inside the original model. Let us illustrate the general method by explaining how to interpret the theory ZFC + $\neg$CH in ZF. In ZF, we may define $L$, and thereby define a model of ZFC with a definable global well-order. Inside that model, consider the forcing notion $\text{Add}(\omega, \omega_1)$ to add $\omega_2$ many Cohen reals. Let $B$ be the Boolean completion of this forcing notion in $L$, and let $U \subseteq B$ be the $L$-least ultrafilter on this Boolean algebra. Using the Boolean ultrapower method, we define the quotient structure $L^B/U$ on the class of $B$-names by the relations:

$$
\sigma =_U \tau \iff [\sigma = \tau] \in U; \\
\sigma \in_U \tau \iff [\sigma \in \tau] \in U.
$$

The model $L^B/U$ consists of the $=_U$ equivalence classes of $B$-names in $L$ under the membership relation $\in_U$. The Boolean ultrapower Łoś theorem shows that $L^B/U$ satisfies every statement $\varphi$ whose Boolean value $[\varphi]$ is in $U$, and so this is a model of ZFC + $\neg$CH. We should like to emphasize that there is no need in the Boolean ultrapower construction for the ultrafilter $U$ to be generic in any sense, and $U \in L$ is completely fine (see extensive explanation in [HS06, §2]). By using the $L$-least such ultrafilter $U$ on $B$, therefore, the Boolean ultrapower quotient $L^B/U$ is a definable interpreted model of ZFC + $\neg$CH inside the original model, as desired.

Essentially similar methods work with the other forcing cases. Any forceable theory will hold in an interpreted model, using the forcing notion and the ultrafilter $U$ in the original model as parameters; when forcing over $L$ or HOD one may use the corresponding definable well-order to define suitable parameters, as we did above. \qed
Apter, Gitman and Hamkins [AGH12] similarly establish in numerous instances that one can find transitive inner models $\langle M, \in \rangle$ of certain large cardinal theories usually obtained by forcing. For example, if there is a supercompact cardinal, then there is a definable transitive inner model with a Laver indestructible supercompact cardinal, and another inner model with a non-indestructible supercompact cardinal, and another with a strongly compact cardinal that was also the least measurable cardinal, and so on.

As we emphasized in the introduction of this article, theorem 7 is concerned with mutual interpretation rather than bi-interpretation. The interpretation methods used in this theorem are not invertible, and when following an interpretation, one cannot in general get back to the original model. Every interpretation of one set theory in another will inevitably involve a loss of set-theoretic information in the models in which it is undertaken. The main lesson of Enayat’s theorem (theorem 12) is that this problem is unavoidable: in fact, none of the theories above are bi-interpretable and there is no nontrivial bi-interpretation phenomenon to be found in set theory at the strength of ZF.

5. Further background on interpreting in models of set theory

Before getting to the main results, let us prove some further useful background material on interpretations in models of set theory. We have already mentioned in the remarks after definition 2 that in models of set theory, we do not need to use $k$-tuples in interpretation, since we have definable pairing functions. But also, it turns out that when interpreting in a model of ZF, we do not need the equivalence relation:

**Theorem 8.** If a structure $A$ is interpreted in a model of ZF set theory, then there is an interpretation in which the equivalence relation is the identity.

**Proof.** The argument amounts to what is known as ‘Scott’s trick.’ If $\approx$ is a definable equivalence relation in a model $M \models ZF$, then let us replace every equivalence class $[a]_\approx$ with the Scott class $[a]_\approx^*$, which consists of the elements of the equivalence class having minimal possible $\in$-rank in that class. This is a definable set in $M$, and from the Scott class one can identify the original $\approx$ equivalence class. So we can replace the original interpretation modulo $\approx$ with the induced interpretation defined on the Scott classes. And since two Scott classes correspond to the same interpreted object if and only if they are equal, we have thereby eliminated the need for the equivalence relation. \qed

We are unsure whether this theorem holds in the case of ZFC$^-$, that is, in set theory without the power set axiom. In ZFC$^-$, after all, the minimal rank instances from a class may still form a proper class, and so Scott’s trick doesn’t succeed in reducing the class to a set.

**Question 9.** Is there a structure that is interpretable in a model of ZFC$^-$, but only by means of a nontrivial equivalence relation?

Next, we establish the phenomenon of automatic bi-interpretability for well-founded models of set theory.
**Theorem 10.** If a well-founded model $M$ of $\text{ZF}^-$ is interpreted in itself via $i : M \to \bar{M}/\simeq$, then the interpretation isomorphism map $i$ is unique and furthermore, it is definable in $M$.

**Proof.** Assume that $M$ is a well-founded model of $\text{ZF}^-$, which we may assume without loss to be transitive, and that $\langle M, \in \rangle$ is interpreted in itself via the interpretation map $i : \langle M, \in \rangle \cong \langle \bar{M}, \bar{\in} \rangle / \simeq$, where $\bar{M} \subseteq M$ is a definable class in $M$ with a definable relation $\bar{\in}$, and $\simeq$ is a definable equivalence relation on $\bar{M}$, which is a congruence with respect to $\bar{\in}$.

Since models of set theory admit pair functions, however, we may assume directly that $\bar{M} \subseteq M$ rather than $\bar{M} \subseteq M^k$ with $k$-tuples.

Let $\pi = i^{-1} : \bar{M} \to M$ be the inverse map, viewed as a map on $\bar{M}$. Since $i$ is an isomorphism, it follows that $b \in a \iff i(b) \bar{\in} i(a)$, and consequently, $\pi(\bar{a}) = \{ \pi(b) \mid b \bar{\in} \bar{a} \}$. Thus, $\pi$ is precisely the Mostowski collapse of the relation $\bar{\in}$ on $\bar{M}$, which is well-founded precisely because it has a quotient that is isomorphic to $\langle M, \in \rangle$. Thus, the map $i$ is unique, for it is precisely the inverse of the Mostowski collapse of $\langle \bar{M}, \bar{\in} \rangle / \simeq$, as computed externally to $M$ in the context where $i$ exists.

What remains is to show that $M$ itself can undertake this Mostowski collapse. While the $\simeq$-quotient of $\bar{\in}$ is a well-founded extensional relation, the subtle issue is that $\bar{\in}$ may not literally be set-like, since the $\simeq$-equivalence classes themselves may be proper classes, and in $\text{ZF}^-$ we are not necessarily able to pick representatives globally from the equivalence classes. So it may not be immediately clear that $M$ can undertake the Mostowski collapse. But, we show, it can.

In $M$, let $I$ be the class of pairs $(a, \bar{a})$ where $a \in M$ and $\bar{a} \in \bar{M}$ and there is a set $A \in M$ with (i) $A \subseteq M$; (ii) $A$ is transitive with respect to $\bar{\in}$ modulo $\simeq$, in the sense that if $x \bar{\in} y \in A$, then there is some $x' \in A$ such that $x \simeq x'$; and (iii) $\langle A, \bar{\in} \rangle / \simeq$ is isomorphic to $\langle \text{TC} \{ \{a\} \}, \in \rangle$. Note that if $(a, \bar{a})$ are like this, then the isomorphism will be exactly the inverse of the Mostowski collapse of $\langle A, \bar{\in} \rangle / \simeq$.

We claim by $\epsilon^{\bar{M}}$-induction (externally to $M$) that every $\bar{a} \in \bar{M}$ is associated in this way with some $a \in M$, and furthermore, the association $I$ is simply the map $i : a \mapsto \bar{a}$. To see this, suppose $\bar{a} \in \bar{M}$ and every $\epsilon^{\bar{M}}$-element of $\bar{a}$ is associated via $I$. Since $i$ is an isomorphism, there is some $a$ such that $i(a) = \bar{a}$. Inductively, every $\bar{b} \in \epsilon^{\bar{M}} \bar{a}$ is associated with some $b$ for which $i(b) = \bar{b}$, and all such $b$ are necessarily elements of $a$. So $M$ can see that every $\bar{b} \in \epsilon^{\bar{M}} \bar{a}$ is associated by the class $I$ with a unique element $b \in a$. For each $b$, there are various sets $B_b$, with largest element $\bar{b}$ realizing that $(b, \bar{b}) \in I$. By the collection axiom, we can find a single set $B$ containing such a set for every $b \in a$, serving as a single witnessing set. Let $A = \{ \bar{a} \} \cup B$, where $\bar{a}$ is a representative of the $\simeq$-class of $i(a)$. This is a set in $M$, it is contained in $\bar{M}$ and it is transitive with respect to $\bar{\in}$ modulo...
\[ \simeq. \text{ Since } \mathcal{B}_b \subseteq \mathcal{A}, \text{ the Mostowski collapse of } \langle \mathcal{A}, \in \rangle / \simeq \text{ agrees with the collapse of } \langle \mathcal{B}_b, \in \rangle / \simeq, \text{ and so sends } b \text{ to } b, \text{ for every } b \in a. \text{ So this set } \mathcal{A} \text{ witnesses that } (a, \bar{a}) \in I, \text{ and agrees with } i \text{ on } a. \text{ So the map } i \text{ is definable in } M, \text{ as desired.} \quad \square \]

Notice that theorem 10 is not true if one drops the well-foundedness assumption. For example, there are models \( M \models \text{ZF}^- \) with nontrivial automorphisms \( i : M \to M \), and any such map is an interpretation of \( M \) in itself (as itself), but such an automorphism can never be definable. Similarly, if ZFC is consistent, then there is a countable computably saturated model \( \langle M, \in^M \rangle \models \text{ZFC} \), and such a model contains an isomorphic copy of itself as an element \( m \); see [GH10, lemma 7]. This provides an interpretation of \( M \) in itself, using \( m \) as a parameter, but the isomorphism cannot be definable nor even amenable to \( M \), since \( M \) cannot allow a definable injection from the universe to a set.

**Corollary 11.** *Every instance of mutual interpretation amongst well-founded models of ZF\(^-\) is a bi-interpretation. Indeed, if \( M \) is a well-founded model of ZF\(^-\) and mutually interpreted with any structure \( N \) of any theory, as in the figure below, then the isomorphism \( i : M \to \overline{M} \) is definable in \( M \).*

**Proof.** This follows immediately from theorem 10, since any instance of mutual interpretation leads to an instance of a model being interpreted inside itself, and so by theorem 10 the interpretation maps will be definable. \( \square \)

We are unsure whether we can achieve the half-way situation, where a transitive model \( M \models \text{ZF}^- \) is mutually interpreted with some nonstandard model \( N \models \text{ZF}^- \), without being bi-interpreted.

**6. Bi-interpretation does not occur in models of ZF**

Finally, we come to the heart of the paper, concerning the extent of bi-interpretation in set theory. Let us begin with Enayat’s theorem, showing that distinct models of ZF are never bi-interpretable. A theory \( T \) is *solid*, if whenever two models of \( T \) are bi-interpretable, then they are isomorphic.

**Theorem 12** (Enayat [Ena16]). *ZF is solid. That is, if two models of ZF are bi-interpretable, then they are isomorphic.*

**Proof.** Let \( M \) and \( N \) be bi-interpretable models of ZF, where \( N \) is definable in \( M \) and \( \overline{N} \) is definable in \( N \) and \( N \) in \( \overline{M} \) and \( i \) is an isomorphism of \( M \) with \( \overline{M} \) that is definable in \( M \), and \( j \) is an isomorphism of \( N \) with \( \overline{N} \) that is definable in \( N \), as illustrated in the figure below. Suppose that \( N \) thinks \( A \) is a nonempty subclass of \( \overline{M} \). Consider the pre-image \( A' = \{ i^{-1}(x) \mid x \in^N A \} \), which is a nonempty class in \( M \). So it must have an \( \in^M \)-minimal element \( a \in^M A' \). It follows that \( i(a) \)
is an $\in^M$-minimal element of $A$, and so $N$ sees $\in^M$ as well founded. Because of this fact, the ordinals of the models will be comparable, and they will have some ordinals isomorphically in common. We claim that for these common ordinals, the rank-initial segments of $M$, $N$ and $\overline{M}$ are isomorphic. Specifically, if $\alpha$ is an ordinal in $M$, which is isomorphic to the ordinal $\alpha^* = i(\alpha)$ in $\overline{M}$, and this happens to be isomorphic in $N$ to an ordinal $\alpha^*$ in $N$, then we claim that $N$ can see that $\langle V_{\alpha^*}, \in \rangle^N$ is isomorphic to $\langle V_{\alpha}, \in \rangle^M$, which we know is isomorphic to $\langle V_{\alpha}, \in \rangle^M$. This is certainly true when $\alpha = 0$, and if it is true at $\alpha$, then it will also be true at $\alpha + 1$, because every subset of $V_{\alpha}^N$ exists also as a class in $M$ and therefore can be pushed via $i$ to the corresponding subset of $V_{\alpha}^M$; and conversely, every subset of $V_{\alpha}^M$ in $\overline{M}$ exists also as a class in $M$ and can be pulled back by the isomorphism with $V_{\alpha^*}$. So in $N$ we may extend the isomorphism canonically from $V_{\alpha}^N \cong V_{\alpha}^M$ to $V_{\alpha^*}^N \cong V_{\alpha^*}^M$. Furthermore, because $V_{\alpha}^N$ is transitive in $N$ and transitive sets are rigid, the isomorphisms must be unique, and so at limit stages the isomorphism will simply be the coherent union of the isomorphisms at the earlier stages.

\[
\begin{array}{c}
\text{M} \\
\text{N}
\end{array}
\]

If the ordinals of $N$ run out before the ordinals of $M$, then $\text{Ord}^N$ is isomorphic to an ordinal $\lambda$ in $M$ and hence to $i(\lambda) = \overline{\lambda}$ in $\overline{M}$. In this case, the isomorphism from the previous paragraph will mean that $N = V_{\text{Ord}}^N \cong V_{\overline{\lambda}}^M$, which is isomorphic to $V_{\lambda}^M$. So $M$ will see that $N$ is bijective with a set. But $M$ has a bijection of itself with $\overline{M}$, which is contained in $N$. So $M$ will think that the universe is bijective with a set, which is impossible in ZF. If in contrast the ordinals of $M$ and hence $\overline{M}$ run out before the ordinals of $N$, then $N$ will similarly look upon $\overline{M}$ as a set, and yet, $N$ has a bijection $j$ of itself to $\overline{N}$ inside $\overline{M}$, which would again contradict ZF. So the ordinals of the three models must be exactly isomorphic, and so the isomorphisms of the rank-initial segments of the models in fact produce an isomorphism of $N = V_{\text{Ord}}^N$ with $\overline{M} = V_{\text{Ord}}^M$ and hence with $M$, as desired.

Albert Visser had initially proved a corresponding result in 2004 for Peano arithmetic PA in [Vis04, see pp.52–55]. Ali Enayat had followed with a nice model-theoretic argument showing specifically that ZF and ZFC are not bi-interpretable, using the fact that ZFC models can have no involutions in their automorphism groups, but ZF models can, before finally proving the general version of his theorem, for ZF, for second-order arithmetic $Z_2$ and for second-order set theory KM in [Ena16]. The ZF version was apparently also observed independently by Harvey Friedman and Albert Visser, by Fedor Pakhomov, as well as by ourselves [Ham18], before we had realized that this was a rediscovery.
Enayat defines that a theory $T$ is tight, if whenever two extensions of $T$ are bi-interpretable, then they are the same theory.

**Corollary 13.** $ZF$ is tight. That is, no two distinct set theories extending $ZF$ are bi-interpretable.

This corollary follows from theorem 12, because in fact every solid theory is tight. If $T$ is solid, and $T_1, T_2$ are distinct bi-interpretable extensions of $T$, then consider any model $M \models T_1$. By the bi-interpretability of the theories, $M$ is bi-interpretable with a model $N \models T_2$. Since these are both models of $T$, it follows by the solidity of $T$ that $M$ and $N$ are isomorphic, and so $M$ is a model of $T_2$. Since the argument also works conversely, the two theories are the same.

In particular, $ZF$ is not bi-interpretable with $ZFC$, nor with $ZFC+CH$, nor $ZFC+\neg CH$ and so on. Theorem 7 cannot be strengthened from mutual interpretation to bi-interpretation, and there is no nontrivial bi-interpretation phenomenon in set theory amongst the models or theories strengthening $ZF$.

Furthermore, we claim that there is no mutual interpretation phenomenon amongst the well-founded models of $ZF$.

**Theorem 14.** Nonisomorphic well-founded models of $ZF$ are never mutually interpretable.

**Proof.** Corollary 11 shows that every instance of mutual interpretation amongst the well-founded models of $ZF$ is a bi-interpretation, but theorem 12 shows that bi-interpretation amongst models of $ZF$ occurs only between isomorphic models. □

Let us now explain how to deduce the solidity of $ZF$ by means of the following ‘internal categoricity’ theorem of Jouko Väänänen. He had stated the result in [Vä19] for $ZFC$, but his argument did not use AC, and so we state and prove it here for $ZF$.

**Theorem 15** (Väänänen [Vä19]). Assume that $\langle V, \in, \in \rangle$ is a model of $ZF$ with respect to both membership relations $\in$ and $\in$, in the common language. More precisely, it is a model of $ZF_{\in}(\in)$, using $\in$ as the membership relation and $\in$ as a class predicate and also of $ZF_{\in}(\in)$, using $\in$ as the membership relation and $\in$ as a class predicate. Then $\langle V, \in \rangle \cong \langle V, \in \rangle$, and furthermore, there is a unique definable isomorphism in $\langle V, \in, \in \rangle$.

**Proof.** Let us begin by observing that by the foundation axiom, every nonempty definable $\in$-class will admit an $\in$-minimal element, and similarly every nonempty definable $\in$-class will have an $\in$-minimal element. Thus, both $\in$ and $\in$ will be well-founded relations, regardless of whether we use $\in$ or $\in$ as the membership relation. In particular, the orders $\langle \text{Ord}(V, \in), \in \rangle$ and $\langle \text{Ord}(V, \in), \in \rangle$ will be comparable well-ordered classes, and we may assume without loss that the former is isomorphic to an initial segment of the latter. Thus, for every $\in$-ordinal $\alpha$, there is an $\in$-ordinal $\bar{\alpha}$ to which it is isomorphic, and furthermore this ordinal is unique, and the isomorphism is unique. We claim that $\langle V_\alpha, \in \rangle \cong \langle V_\in, \in \rangle$, by induction on $\alpha$. Since these are transitive sets and hence rigid, the isomorphism is unique. The claim is clearly true for $\alpha = 0$. It remains true through limit ordinals, since the isomorphism in that case is simply the union of the previous isomorphisms. At successor stages, we need only observe that any isomorphism of $\langle V_\alpha, \in \rangle$ with $\langle V_\in, \in \rangle$ extends to an isomorphism of the power set, since any $\in$-subset of $V_\alpha$ transfers to an $\in$-subset
of $V, \pi$ by pointwise application of the isomorphism, and vice versa. If $\text{Ord}^{(V, \in)}$ is taller than $\text{Ord}^{(V, \bar{\in})}$, then we can see that $(V, \in) \cong (\bar{V}, \bar{\in})$, where $\gamma \in \text{Ord}^{(\bar{V}, \bar{\in})}$ is the first $\bar{\in}$-ordinal not arising from an $\in$-ordinal. But now $\bar{V}_{\gamma+1}$, which is the $\bar{\in}$-power set of $\bar{V}_\gamma$, is a subclass of $V$ and therefore injects into $\bar{V}_\gamma$, contrary to Cantor’s theorem. So the ordinals are the same height, and thus we have achieved $(V, \in) \cong (V, \bar{\in})$. This isomorphism is definable, since every $V_\alpha$ is isomorphic to $\bar{V}_\alpha$ by a unique isomorphism.

□

Theorem 12, we claim, can be seen as a consequence of this theorem—and indeed the proofs we gave are roughly analogous—because if two models of ZF are bi-interpretable, then by the results mentioned in section 5, we do not need tuples or nonidentity equivalence relations. So the interpretations will interpret each model as a subclass of the other. By the class version of the Shröder-Bernstein theorem, these injections can be transformed into bijection of the models, and in this way the bi-interpretation can be transformed into a bi-interpretation synonymy (see also [FV14]). Thus, we reduce to the case of two membership relations $(M, \in, \bar{\in})$ on the whole universe, with each relation being definable when using the other as the membership relation. Because of this, the model will satisfy the theory $ZF(\in, \bar{\in})$, and consequently the models will be isomorphic $(M, \in) \cong (M, \bar{\in})$ by Väänänen’s theorem.

7. Väänänen internal categoricity fails for ZFC$^-$

Theorems 12 and 15 were both proved with arguments that made a fundamental use of the $V_\alpha$ hierarchy, thereby relying on the power set axiom. Was the use of the $V_\alpha$ hierarchy significant? Can one prove the analogues of the theorems for ZFC$^-$, that is, for set theory without the power set axiom? (See [GHJ16] for a subtlety about exactly what the theory ZFC$^-$ is; one should include the collection axiom, and not just replacement.) The answer is negative. We shall prove that neither theorem is true for ZFC$^-$, not even in the case of well-founded models. Let us begin by proving that the Väänänen internal categoricity theorem fails for ZFC$^-$. 

**Theorem 16.** There is a transitive set $M$ and an alternative well-founded membership relation $\bar{\in}$ on $M$, such that $(M, \in, \bar{\in})$ satisfies ZFC$^-(\in, \bar{\in})$, that is, ZFC$^-$ using either $\in$ or $\bar{\in}$ as the membership relation and allowing both as class predicates, such that $(M, \in)$ is not isomorphic to $(M, \bar{\in})$, although both are well-founded.

*Proof.* Let us argue first merely that the situation is consistent with ZFC. Assume that Luzin’s hypothesis holds, that is, $2^{\omega_1} = 2^{\omega_1}$; for example, this holds after adding $\omega_2$ many Cohen reals over a model of GCH, since in this case we would have $2^{\omega_1} = 2^{\omega_1} = \omega_2$. It follows that $H_{\omega_1}$ and $H_{\omega_2}$ are equinumerous, and so there is a bijection $\pi : H_{\omega_1} \to H_{\omega_2}$. Let $\bar{\in}$ be the image of $\in$ on $H_{\omega_1}$ under this bijection, and let $\bar{\in}$ be the preimage of $\in$ on $H_{\omega_2}$. Thus, $\pi$ is an isomorphism of the structures

$$\pi : (H_{\omega_1}, \in, \bar{\in}) \cong (H_{\omega_2}, \bar{\in}, \in).$$

The first of these is a model of ZFC$^-(\bar{\in})$, using $\in$ as membership, since $H_{\omega_1}$ is a model of ZFC$^-$ with respect to any predicate. Similarly, the second is a model of ZFC$^- (\bar{\in})$, again using $\in$ as membership, since we can likewise augment $H_{\omega_2}$ with any predicate. By following the isomorphism, an equivalent way to say this is that $(H_{\omega_1}, \in, \bar{\in})$ is a model of ZFC$^- (\bar{\in})$, using $\in$ as membership and $\bar{\in}$ as predicate,
and also a model of $\text{ZFC}^-(\in)$, using $\in$ as membership this time and $\in$ as a class predicate. But $\langle H_{\omega_1}, \in \rangle$ is not isomorphic to $\langle H_{\omega_1}, \in \rangle$, because the first thinks every set is countable and the second does not. So this model is just as desired.

To get outright existence of the models, observe first that by taking an elementary substructure, we can find a countable model of the desired form in the forcing extension. Furthermore, the existence of a countable well-founded model $\langle M, \in, \mathcal{E} \rangle$ with the desired properties is a $\Sigma_1^1$ assertion, which is therefore absolute to the forcing extension. And so therefore there must have already been such an example in the ground model, without need for any forcing. 

\section{ZFC$^-$ is neither solid nor tight}

In order to show that ZFC$^-$ is not solid, we need to provide different models of ZFC$^-$ that are bi-interpretable, but not isomorphic. Let us begin by describing merely how this can happen in some models of set theory. For this, we shall employ a more refined version of the argument used to prove theorem 16.

\textbf{Theorem 17.} In the Solovay-Tennenbaum model of $\text{MA}^-\text{–CH}$ obtained by c.c.c. forcing over the constructible universe, the structures $\langle H_{\omega_1}, \in \rangle$ and $\langle H_{\omega_2}, \in \rangle$ are bi-interpretable. Thus, there can be two well-founded models of ZFC$^-$ that are bi-interpretable, but not isomorphic.

The features we require in the Solovay-Tennenbaum model are the following:

- $H_{\omega_1}$ has a definable almost disjoint $\omega_1$-sequence of reals;
- every subset $A \subseteq \omega_1$ is coded by a real via almost-disjoint coding with respect to this sequence.

In other words, there should be a definable $\omega_1$-sequence $\langle a_\alpha \mid \alpha < \omega_1 \rangle$ of infinite subsets $a_\alpha \subseteq \omega$ with any two having finite intersection; and for every $A \subseteq \omega_1$ there should be a $a \subseteq \omega$ for which $\alpha \in A \leftrightarrow a \cap a_\alpha$ is infinite. These properties are true in the Solovay-Tennenbaum model $L[G]$, because we can define the almost disjoint family in $L$, and by Martin's axiom every subset of $\omega_1$ is coded with respect to it by almost-disjoint coding.

\textit{Proof.} Let us assume we have the two properties isolated above. The structure $\langle H_{\omega_1}, \in \rangle$, of course, is a definable substructure of $\langle H_{\omega_2}, \in \rangle$, which gives one direction of the interpretation. It remains to interpret $\langle H_{\omega_2}, \in \rangle$ inside $\langle H_{\omega_1}, \in \rangle$ and to prove that this is a bi-interpretation. The second assumption above implies, of course, that $2^\omega = 2^{\omega_1}$, and so at least the two structures have the same cardinality. But more, because the almost-disjoint sequence is definable in $H_{\omega_1}$ and all the reals are there, the structure $\langle H_{\omega_1}, \in \rangle$ in effect has access to the full power set $P(\omega_1)$; the subsets $A \subseteq \omega_1$ are uniformly definable in the structure $\langle H_{\omega_1}, \in \rangle$ from real parameters. And furthermore, every set in $H_{\omega_2}$ is coded by a subset of $\omega_1$. Specifically, if $x \in H_{\omega_2}$ then $x$ is an element of a transitive set $t$ of size $\omega_1$, with $\langle t, \in \rangle \cong \langle \omega_1, E \rangle$ for some well-founded extensional relation $E$ on $\omega_1$. The set $x$ can be in effect coded by specifying $E$ and the ordinal $\alpha$ representing $x$ in the structure $\langle \omega_1, E \rangle$. Since $H_{\omega_1}$ contains all countable sequences of ordinals, it can correctly identify which relations $E$ on $\omega_1$, as encoded by a real via the almost-disjoint coding, are well-founded and extensional. Let us define an equivalence relation $a \simeq a'$ for reals $a, a' \subseteq \omega$ coding such pairs $(\alpha, E)$ and $(\alpha', E')$ that represent the same set, which happens just in case there is an isomorphism of the hereditary $E$-predecessors of $\alpha$ to the
Theorem 19. If there is no projectively definable somewhat the need in theorem 17 for assuming that we were working close to being interesting in connection with theorem 17. The following theorem explains for some ⟨y⟩ interpreted copy of a Hereditary E-sequences of distinct reals, such a sequence would be definable (from parameters) in the structure ⟨Hω1, ∈⟩. But this latter structure is interpreted in the reals in second-order arithmetic, by coding hereditarily countable sets by relations.
on $\omega$. By means of this interpretation, assertions in $H_{\omega_1}$ can be viewed as projective assertions. And so there would be a projectively definable $\omega_1$-sequence of distinct reals, contrary to our assumption. \hfill \Box

Meanwhile, we can extend theorem 17 from the consistency result to prove outright in ZFC that there are always transitive counterexample models to be found.

**Theorem 20.** The theory $\text{ZFC}^-$ is not solid, not even for well-founded models. Specifically, there are transitive models $\langle M, \in \rangle$, $\langle N, \in \rangle$ of $\text{ZFC}^-$ that are bi-interpretable, but not isomorphic.

*Proof.* By theorem 17, there are such transitive sets in a forcing extension of $L$. In that model, by taking a countable elementary substructure, one can find countable transitive sets with the desired feature. And furthermore, the assertion that there are such countable transitive sets like that, for which the particular bi-interpretation definitions work, is a statement of complexity $\Sigma^1_2$. By Shoenfield absoluteness, this statement must already be true in $L$, and hence also in $V$. So there are countable transitive models $\langle M, \in \rangle$ and $\langle N, \in \rangle$ of $\text{ZFC}^-$ that are bi-interpretable, but not isomorphic. \hfill \Box

And we can also use the argument to show that $\text{ZFC}^-$ is not tight.

**Theorem 21.** $\text{ZFC}^-$ is not tight.

*Proof.* We shall find two extensions of $\text{ZFC}^-$ that are bi-interpretable, but not the same theory. The basic idea is to let $T_1$ and $T_2$ be the theories describing the situation of $\langle H_{\omega_1}, \in \rangle$ and $\langle H_{\omega_2}, \in \rangle$ in theorem 17. Specifically, let $T_2$ assert $\text{ZFC}^-$ plus the assertion that $\omega_1$ exists but not $\omega_2$, that $\omega_1 = \omega^L_1$, that $\omega_2 = \omega^L_2$, and that every subset of $\omega_1$ is coded by a real using the almost-disjoint coding with respect to the $L$-least almost-disjoint family $\langle a_\alpha \mid \alpha < \omega_1 \rangle$. Let $T_1$ be the theory $\text{ZFC}^-$ plus the assertion that every set is countable and that the interpretation of $H_{\omega_2}$ in $H_{\omega_1}$ used in the proof of theorem 17 defines a model of $T_2$. That is, $T_1$ asserts that the interpretation does indeed interpret a model of $T_2$. These two theories are bi-interpretable, using the interpretations provided in the proof of theorem 17, because any model of $T_1$ interprets a model of $T_2$, and any model of $T_2$ can define its $H_{\omega_1}$, which will be a model of $T_1$, and the composition maps are definable just as before. \hfill \Box

9. **Zermelo set theory is neither solid nor tight**

We should like now to consider the case of Zermelo set theory. We shall prove that nontrivial instances of bi-interpretation occur in models of Zermelo set theory, and consequently Zermelo set theory is neither solid nor tight. Specifically, we shall prove the following:

**Theorem 22.**

1. $\text{Z}$ is not solid, not even for well-founded models: there are bi-interpretable well-founded models of Zermelo set theory that are not isomorphic.
2. Every model of $\text{ZF}$ is bi-interpretable with a transitive inner model of Zermelo set theory, in which the replacement axiom fails.
3. $\text{Z}$ is not tight: there are distinct bi-interpretable strengthenings of $\text{Z}$. 
Our argument will make fundamental use of the model-construction method of Adrian Mathias [Mat01]. Mathias had used his methods to construct diverse interesting models of Zermelo set theory, some of them quite peculiar, such as a transitive inner model of Zermelo set theory containing all the ordinals, but not having $V_\omega$ as an element. Let us briefly review his method. The central definition is that a class $C$ is fruitful, if

1. every $x \in C$ is transitive;
2. $\text{Ord} \subseteq C$;
3. $x \in C$ and $y \in C$ implies $x \cup y \in C$;
4. $x \in C$ and $y \subseteq P(x)$ implies $x \cup y \in C$.

Fruitful classes, Mathias proves, lead to transitive models of Zermelo set theory as follows.

**Theorem 23** (Mathias [Mat01, prop. 1.2]). If $C$ is fruitful, then $M = \bigcup C$ is a supertransitive model of Zermelo set theory with the foundation axiom.

A transitive set $M$ is supertransitive, if it contains every subset of its elements, that is, if $x \subseteq y \in M \rightarrow x \in M$. In the context of Zermelo set theory, we take the foundation axiom in the form of the $\in$-induction scheme, which is equivalent over Z to the ‘transitive containment’ assertion that every set is an element of a transitive set.

The importance of the theorem is that Mathias explains how to construct a great variety of fruitful classes $T^{Q,G}$, defined in terms of a function $Q : \omega \rightarrow V_\omega$ and a family $G$ of functions from $\omega$ to $\omega$, which specify rates of growth for the allowed sets. Namely, a transitive $x$ is in $T^{Q,G}$ just in case its rate-of-growth function

\[ f^Q_x(n) = |x \cap Q(n)| \]

is in $G$. Under general hypotheses, he proves, $T^{Q,G}$ is fruitful.

For our application, it will suffice to consider one of his key examples, described in [Mat01, §3]. Namely, we consider the case of $T^{R,F}$, where $R(n) = V_n$ is the rank-initial segment function and $F$ is the collection of functions $f$ bounded by one of the superexponential functions $b_k$ of the form

\[ b_k : n \mapsto 2^{2^n} \]

for some $k \in \omega$. Mathias proves in this instance that $T^{R,F}$ is fruitful, and therefore by theorem 23 the class $M = \bigcup T^{R,F}$ is a supertransitive model of Zermelo set theory. The elements of $M$ are precisely those sets $x$ whose transitive closure has a rate of growth in $V_n$ bounded by some $b_k$,

\[ x \in M \iff \exists k \forall n \ |TC(\{x\}) \cap V_n| \leq b_k(n). \]

Since each $b_k$ has a fixed superexponential depth $k$, it follows that each of them has strictly slower asymptotic growth than the full tetration function

\[ n \mapsto |V_n| = 2^{2^n} \]

and therefore the set $V_\omega$ is not in $M$ because it grows too quickly, [Mat01, thm. 3.8].

In order to prove the claims of theorem 22, we shall show that the original ZF model $\langle V, \in \rangle$ is bi-interpretable with the Zermelo model $\langle M, \in \rangle$ defined by $M = \bigcup T^{R,F}$ above. We already have one direction of the interpretation, since we have defined $M$ as a transitive inner model inside $V$. The difficult part will be conversely
to provide an interpretation of $\langle V, \in \rangle$ inside $\langle M, \in \rangle$. For this, we shall make use of the Zermelo tower hierarchy, defined for any set $a$ in $V$ as follows:

1. $V^{(a)}_0 = \emptyset$;
2. $V^{(a)}_{\alpha+1} = \{a\} \cup (P(V^{(a)}_\alpha) - \{\emptyset\})$, for successor ordinals;
3. $V^{(a)}_\lambda = \bigcup_{\alpha<\lambda} V^{(a)}_\alpha$, for limit ordinals.

Mathias had introduced the Zermelo tower idea in [Mat01, §4], but defined and considered it only at finite stages and $\omega$, whereas we continue the construction, crucially, through all the ordinals. Our reason for doing so is that the full Zermelo tower $V^{(a)}$ is a definable copy of the original universe, as we prove here:

**Lemma 24.** For any set $a$, the universe $\langle V, \in \rangle$ is definably isomorphic to the full Zermelo tower $\langle V^{(a)}, \in \rangle$.

**Proof.** The isomorphism from $V$ to $V^{(a)}$ is simply the operation replacing every hereditary instance of $\emptyset$ in a set with $a$. Specifically, we define the replacing operation $x \mapsto x^{(a)}$ as follows:

- $\emptyset^{(a)} = a$.
- $x^{(a)} = \{y^{(a)} \mid y \in x\}$

A simple argument by transfinite induction now shows that this is an isomorphism of every $V^\alpha$ with $V^{(a)}_\alpha$, which establishes the theorem. \(\square\)

Next, we show that for suitable sets $a$, the Zermelo tower is contained within the Zermelo universe $M$ we defined above.

**Lemma 25.** If $a$ is any infinite transitive set in $M$, then $V^{(a)} \subseteq M$.

**Proof.** Suppose that $a$ is an infinite transitive set in $M$. It therefore obeys the growth rate requirement, and so there is some $k$ for which $|a \cap V_n| \leq b_k(n)$ for every $n$. Notice that $a \cup V^{(a)}_n$ is a transitive set, and since (i) $u^{(a)}$ is infinite for every set $u$, and (ii) every element of $V_n$ is hereditarily finite, it follows that $a \cap V_n \subseteq V^{(a)}_n \cap V_n = a \cap V_n$.

Consequently, $a \cup V^{(a)}_\alpha$ obeys the growth-rate requirement, and consequently every $V^{(a)}_\alpha$ is in $M$. So $V^{(a)} \subseteq M$. \(\square\)

**Lemma 26.** For any set $a \in M$, the class $V^{(a)}$ is definable in $M$ from parameter $a$.

**Proof.** One might naturally want to define $V^{(a)}$ in $M$ by transfinite recursion, as we did in defining the hierarchy $V^{(a)}_\alpha$ above. The problem with this is that $M$ is a model merely of Zermelo set theory, in which such definitions by transfinite recursion do not necessarily succeed. Nevertheless, in instances as here where we know independently that the recursion does have a solution in $M$, the recursive definition does in fact succeed in defining it.

But instead of that definition, let us consider another simple definition. Define that a *terminal* $\in$-descent from a set $x$ is a finite $\in$-descending sequence of sets from $x$ to $\emptyset$:

$$x = x_n \ni x_{n-1} \ni \cdots \ni x_1 \ni x_0 = \emptyset.$$  

We claim that $V^{(a)}$ consists in $M$ precisely of those sets $x$ for which every terminal $\in$-descent passes through the set $a$, in the sense that $a = x_i$ for some $i$. Certainly
every $x \in V^{(a)}$ is like this, since if $y \in x \in V^{(a)}_{\alpha+1}$, then either $x = a$ or $y \in V^{(a)}_{\alpha}$, and by induction the terminal descents from $y$ will pass through $a$. Conversely, assume that every terminal $\varepsilon$-descent from $x$ passes through $a$. So either $x = a$, in which case it is in $V^{(a)}$, or else the assumption about terminal descents is also true of the elements of $x$. By $\varepsilon$-induction, therefore, the elements of $x$ are all in $V^{(a)}$, and this implies that $x$ is in $V^{(a)}_{\gamma}$ at an ordinal stage $\gamma$ beyond the stages where the elements of $x$ are placed into $V^{(a)}$. □

If we take $a = \omega$, or some other definable infinite transitive set in $M$, then we don’t need $a$ as a parameter, and so we’ve proved that $\langle V, \varepsilon \rangle$ and $\langle M, \varepsilon \rangle$ are at least mutually interpretable; each has a definable copy in the other.

**Lemma 27.** The mutual interpretation of $\langle V, \varepsilon \rangle$ and $\langle M, \varepsilon \rangle$ is a bi-interpretation.

**Proof.** We’ve seen already that $M$ is a definable transitive inner model of $V$, and $\langle V, \varepsilon \rangle$ is definably isomorphic to its copy $\langle V^{(a)}, \varepsilon \rangle$ in $M$, isomorphic by the map $x \mapsto x^{(a)}$, which is definable in $V$, assuming as we have that we use $a = \omega$ or some other definable infinite transitive set in $M$.

Conversely, we’ve seen that $M$ can define the model $\langle V^{(a)}, \varepsilon \rangle$. Inside that model, we define the copy $M^{(a)}$ of $M$, just as $M$ is definable in $V$. For this to be a bi-interpretation, we need to show that the isomorphism of $M$ with $M^{(a)}$ is definable in $M$. This isomorphism is precisely the function $i : x \mapsto x^{(a)}$, applied to $x \in M$. The reason is that the transitive closure of $i \upharpoonright x = \{ (y, y^{(a)}) \mid y \in x \}$ can be seen to obey the growth-rate requirement, as $x$ does with its transitive closure and the sets $y^{(a)}$ are either $x$ or have no elements in any $V_{\alpha}$, as in lemma 25 above. Thus, $M$ can see the isomorphism $i \upharpoonright x$ of $\langle x, \varepsilon \rangle$ with $\langle x^{(a)}, \varepsilon \rangle$ for any set $x \in M$. If $x$ is transitive, then since transitive sets are rigid, this isomorphism is unique, and since different transitive sets are never isomorphic, $(x, \varepsilon)$ will not be isomorphic to any $(y, \varepsilon)$ that $M^{(a)}$ thinks is transitive. So $M$ can define $i \upharpoonright x$ for transitive sets $x$ by the assertions that it is an isomorphism of $\langle x, \varepsilon \rangle$ with a set $(y, \varepsilon)$ for a set $y$ that $M^{(a)}$ thinks is transitive. These maps all agree with and union up to the full isomorphism $x \mapsto x^{(a)}$ of $M$ with $M^{(a)}$, which is therefore definable in $M$. So we have a bi-interpretation of $\langle V, \varepsilon \rangle$ with $\langle M, \varepsilon \rangle$. □

Let us now put all this together and explain how we have proved theorem 22.

**Proof of theorem 22.** The argument we have given shows that every model $V \models ZF$ is bi-interpretable with a transitive inner model $M$ of Zermelo set theory, which is not a model of ZF, because it does not even have $V_{\omega}$ as an element. This establishes statement (2) of theorem 22.

For statement (1), we can perform the construction inside a well-founded model of some sufficient fragment of ZF. For example, let us undertake the entire construction inside some $V_{\lambda}$, satisfying $\Sigma_2$-collection, say. We thereby get a transitive model $M_{\lambda} \subseteq V_{\lambda}$ of Zermelo set theory, without $V_{\omega}$, but with $\langle V_{\lambda}, \varepsilon \rangle$ and $\langle M_{\lambda}, \varepsilon \rangle$ bi-interpretable. So there are well-founded models of Zermelo set theory that are bi-interpretable, but not isomorphic.

Finally, let us consider statement (3). We consider the two extensions of Zermelo set theory describing the situation that enabled us to make the main example for
statement (2). That is, the first theory is ZF itself; and the second theory, which we denote ZM, asserts Zermelo set theory Z, plus the assertion that the class $V^{(\omega)}$, defined as in our main argument, is a model of ZF, and the assertion that $M$ is isomorphic to $M^{(\omega)}$ inside that class by the isomorphism definition we provided. These two theories are different, but the argument that we gave above for the models shows that they are bi-interpretable.

Let us close this section by mentioning that the construction is quite malleable with respect to the instance of replacement we want to eliminate. Alternative versions of the construction will construct supertransitive inner models $M$ that satisfy Zermelo set theory, and which have $V_\alpha$ as a set for every ordinal $\alpha < \lambda$, but do not have $V_\lambda$ as a set.

**Theorem 28.** For every limit ordinal $\lambda$, the universe $\langle V, \in \rangle$ is bi-interpretable with a transitive inner model $\langle M, \in \rangle$ satisfying Zermelo set theory with foundation, such that $V_\alpha \in M$ for every $\alpha < \lambda$, but $V_\lambda \notin M$.

**Proof.** We use $M = \bigcup T^{Q,G}$, where $Q : \lambda \to V_\lambda$ is $Q(\alpha) = V_\alpha$, and $G$ consists of the functions $f$ that are bounded by one of the functions $b_\beta$ defined by

1. $b_0(\alpha) = \alpha$;
2. $b_{\beta+1}(\alpha) = 2^{b_\beta(\alpha)}$;
3. if $\eta$ is a limit ordinal, then $b_\eta(\alpha) = \sup_{\beta < \eta} b_\beta(\alpha)$.

This is defined with cardinal arithmetic, not ordinal arithmetic. The class $T^{Q,G}$ consists of all transitive sets $x$ whose rate-of-growth functions

$$f^Q_x(\alpha) = |x \cap V_\alpha|$$

is bounded by one of the functions $b_\beta$. It follows that $T^{Q,G}$ is a fruitful class, and the resulting Zermelo model $M = \bigcup T^{Q,G}$ is a transitive inner model of Zermelo set theory with foundation (in the form of the $\in$-induction scheme). Every $V_\alpha$ for $\alpha < \lambda$ is in $M$, with growth rate bounded by $b_\alpha$, but $V_\lambda$ is not in $M$, since this growth rate exceeds any given $b_\beta$ for fixed $\beta < \lambda$. And $\langle V, \in \rangle$ is bi-interpretable with $\langle M, \in \rangle$ by analogous arguments to the above. \qed

One can use this idea to produce many non-tight theories extending Zermelo set theory. For example, for any theory true in the model $M$ that we produced in the proof of theorem 22 will be bi-interpretable with ZF. For example, that model $M$ satisfies the principle that every definable function from a set to the ordinals is bounded, which can be seen as a weak form of replacement not provable in Z, but true in $M$ precisely because it is a transitive inner model (with all the ordinals) of a model of ZF. In theorem 28 we similarly arranged that $V_\alpha$ exists for all $\alpha$ up to some large definable limit ordinal $\lambda$, which is itself another instance of replacement, and so this gives another extension of Zermelo set theory that is bi-interpretable with ZF.

Let us conclude the paper by distinguishing the theory form of bi-interpretation from a natural model-by-model form of bi-interpretation that might hold for the models of the theory. Specifically, we define that

**Definition 29.** Theories $T_1$ and $T_2$ are **model-by-model** bi-interpretable if every model of one of the theories is bi-interpretable with a model of the other.
The definition in effect drops the uniformity requirement for bi-interpretability of theories, since it could be a different interpretation that works in one model than in another, with perhaps no uniform definition that works in all models of the theory. This is precisely what we should like now to prove.

**Theorem 30.** There are theories $T_1$ and $T_2$ that are model-by-model bi-interpretable, but not bi-interpretable.

**Proof.** Consider the theories

1. $T_1 = \text{ZF}.$
2. $T_2 = \{ \alpha \lor \beta \mid \alpha \in \text{ZF} \land \beta \in \text{ZM} \}.$ (That is, $T_2 = \text{ZM} \lor \text{ZF}$)

where ZM is the theory used in the proof of theorem 22. Observe that $M \models T_2$ if and only if $M \models \text{ZF}$ or $M \models \text{ZM}$. The reverse implication is immediate, and if the latter fails, then $M \not\models \alpha$ and $M \not\models \beta$ for some $\alpha \in \text{ZF}$ and $\beta \in \text{ZM}$, which would mean $M \not\models \alpha \lor \beta$, violating the former.

Now observe simply that every model of ZF is bi-interpretable with itself, of course, and this is a model of $T_2$. And conversely, every model of $T_2$ is either a model of ZF already, or else a model of ZM, which is bi-interpretable with a model of ZF. So these two theories are model-by-model bi-interpretable.

Let us now prove that the theories are not bi-interpretable. Suppose toward contradiction that they are, that $(I, J)$ forms a bi-interpretation, so that $I$ defines a model of $T_1$ inside any model of $T_2$ and $J$ defines a model of $T_2$ inside any model of $T_1$, and the theories prove this and furthermore have definable isomorphisms of the original models with their successive double-interpreted copies. Let $M$ be a model of ZM that is not a model of ZF. Let $I^M$ be the interpreted model of ZF defined inside $M$. This is also a model of ZM, since ZM $\subseteq$ ZF, and so we may use $I$ again to define $I^{I^M}$ to get another model of ZF inside that model. Because we interpreted a model of $T_2$ by $I$, it follows that $I^{I^M}$ and $I^M$ are bi-interpretable as models. But these are both models of ZF, and so by theorem 12 they are isomorphic. By interpreting with $J$ to get the corresponding interpreted ZM models, it follows that $J^M$ and $I^{I^M}$ are isomorphic. Because $(I, J)$ form a bi-interpretation, the first of these is isomorphic to $I^M$, which is a ZF model, while the second is isomorphic to $M$, which is not. This is a contradiction, and so the theories are not bi-interpretable. □

The essence of the proof was that we had a theory ZM that was not solid, but it was a subtheory of a solid theory ZF, with which it was bi-interpretable.

**10. Final remarks**

We have investigated the nature of mutual and bi-interpretation in set theory and the models of set theory. There is surely a vibrant mutual interpretation phenomenon in set theory, with numerous instances of mutual interpretation amongst diverse natural extensions of ZF set theory. This mutual-interpretation phenomenon, however, is less than fully robust in several respects. First, theories extending ZF are never bi-interpretable, and indeed, there cannot be even a single nontrivial instance of models of ZF set theory being bi-interpretable. Worse, amongst the well-founded models of ZF, there cannot even be nontrivial instances of mutual interpretation. The basic lesson of this is that, when one follows the mutual interpretations we mentioned above, there is no getting back. One cannot recover the original model. When you move from a well-founded model of ZFC with large
cardinals to the corresponding determinacy model, you cannot define a copy of the original model again, and the same when interpreting in the other direction. In this sense, interpretation in models of set theory inevitably involves and requires the loss of set-theoretic information.

Meanwhile, we have showed that several natural weak set theories avoid this phenomenon, including Zermelo-Fraenkel set theory $\mathsf{ZFC}^-$ without the power set axiom and Zermelo set theory $\mathsf{Z}$. It remains open just how strong the weak set theories can be, while still admitting non-trivial instances of bi-interpretation. In addition, there remain interesting open questions concerning the circumstances in which $H_{\omega_1}$ and $H_{\omega_2}$ might be bi-interpretable or bi-interpretation synonymous.

References

[AGH12] Arthur W. Apter, Victoria Gitman, and Joel David Hamkins. Inner models with large cardinal features usually obtained by forcing. Archive for Math. Logic, 51:257–283, 2012, 1111.0856. doi:10.1007/s00153-011-0264-5.

[dB65] K. L. de Bouvire. Logical synonymy. Indagationes Mathematicae, 27:622–629, 1965.

[Ena16] Ali Enayat. Variations on a Visserian theme. In Jan van Eijck, Rosalie Iemhoff, and Joost J. Joosten, editors, Liber Amicorum Alberti : a tribute to Albert Visser, pages 99–110. College Publications, London, 03 2016, 1702.07093.

[FF19] Alfredo Roque Freire and Rodrigo de Alvarenga Freire. The ontological import of adding proper classes. Manuscrito, 42(2):85–112, 2019.

[Fre08] Alfredo Roque Freire. On what counts as a translation. In Michal Pelis, editor, The Logica Yearbook 2007. Filosofia, 2008.

[Fre19] Alfredo Roque Freire. Estudio comparado do comprometimento ontológico das teorias de classes e conjuntos. PhD thesis, University of Campinas, 2019.

[FV14] Harvey M Friedman and Albert Visser. When bi-interpretability implies synonymy. Logic Group Preprint Series, 320:1–19, 2014.

[GH10] Victoria Gitman and Joel David Hamkins. A natural model of the multiverse axioms. Notre Dame J. Formal Logic, 51(4):475–484, 2010, 1104.4450. doi:10.1215/00294527-2010-030.

[GHJ16] Victoria Gitman, Joel David Hamkins, and Thomas A. Johnstone. What is the theory $\mathsf{ZFC}$ without Powerset? Math. Logic Q., 62(4-5):391–406, 2016, 1110.2430. doi:10.1002/malq.201500019.

[Ham18] Joel David Hamkins. Different set theories are never bi-interpretable. Mathematics and Philosophy of the Infinite, March 2018. URL http://jdh.hamkins.org/different-set-theories-are-never-bi-interpretable/.

[HS06] Joel David Hamkins and Daniel Seabold. Well-founded Boolean ultrapowers as large cardinal embeddings. pages 1–40, 2006, 1206.6075. URL http://jdh.hamkins.org/boolean-ultrapowers/.

[Mat01] Adrian RD Mathias. Slim models of Zermelo set theory. The Journal of Symbolic Logic, 66(2):487–496, 2001.

[Väi19] Jouko Väänänen. An extension of a theorem of Zermelo. The Bulletin of Symbolic Logic, 25(2):208–212, Mar 2019. doi:10.1017/bsl.2019.15.

[Vis04] Albert Visser. Categories of theories and interpretations. Utrecht University Repository, Logic Group Preprint Series, 228, 2004. URL https://dspace.library.uu.nl/handle/1874/26909.